Understanding The Evolution Of Gene Expression From A Regulatory Network Perspective

by

Bing Yang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Molecular, Cellular and Developmental Biology)
in the University of Michigan
2017

Doctoral Committee:

Professor Patricia Wittkopp, Chair
Associate Professor Scott E. Barolo
Associate Professor Kenneth M. Cadigan
Professor Anuj Kumar
Professor Jianzhi Zhang

ypauling@umich.edu

ORCID: 0000-0001-5972-1521

To Sai, for your love and support

Acknowledgements

I would first thank Patricia Wittkopp for being a great mentor over the 6 years. I am grateful for her support and guidance in pursuing Doctor of Philosophy. I would also thank Fabien Duveau, who has been another mentor for me for the last three years. I would like to thank Andrea Hodgins-Davis, who shows me how to be a good and patient collaborator. I would like to thank Gizem Kalay, who helped me in my first year of Ph.D. and encouraged me to explore the scientific world. I would like to thank Joseph Coolon for providing me great insights on genomic studies. I would like to thank Brian Metzger, who has set up an example for me as what is a good biologist. I would like to thank all other current or alumni members in Wittkopp lab for creating a good environment; Jen, Jon, Alisha, Abby, Kraig, Petra, Jade and all other members. I would like to thank all the committee members: Professor Scott Barolo, Professor Kenneth Cadigan, Professor Anuj Kumar and Professor Jianzhi Zhang, for providing me suggestions on research topics and reading my thesis. I would like to thank all my friends, for the happiness we shared together for the past 6 years in another country. I would like to thank my parents for nothing special, cause they gave me life. Finally, I would like to thank my wife Sai. Without your company, I cannot survive the tough time in my Ph.D. life. This whole Ph.D. thesis is for you.

Table of Contents

List of Figures

List of Tables

Abstract

The evolution of transcriptional regulation has been demonstrated to be a major contributor to phenotypic evolution. One important step in transcriptional regulation is the interaction between transcription factors and their target genes, the organization of which is represented by the Transcriptional Regulatory Network (TRN). Recent studies have shown that structural properties within a TRN provide important information for understanding how different transcriptional patterns are formed in many biological systems. However, it is less clear whether or not those structural properties are also informative in understanding the evolution of transcriptional patterns. In this thesis, I examined the question of whether the number of connections for a gene in a TRN was associated with observed gene expression differences by combining published datasets from multiple related *Drosophila* species. Specifically, I found that increasing number of regulators (in-degree) for a gene was associated with decreasing differences in gene expression and *cis* regulation. Meanwhile, I found no significant relationship between the number of targets (out-degree) for a transcription factor and differences in gene expression. To assess the generality of the conclusions from *Drosophila* species, I inferred a whole-genome transcriptional regulatory network in *Saccharomyces cerevisiae* and combined it with published gene expression datasets involving multiple *Saccharomyces* species to examine the relationship between in-degree/out-degree and differences in gene expression. I found that increasing in-degree was associated with increasing differences in gene expression between two strains of *S. cerevisiae,* but no

significant relationship between in-degree and differences in gene expression was detected in all comparisons between two diverged *Saccharomyces* species. These two studies suggest that whether and how the number of interactions for a gene within a TRN could impact the evolution of the transcription level might depend on the biological system under consideration. Finally, I examined whether and how existing genetic variants that disrupted transcriptional regulation of a yeast gene *TDH3* could influence how random mutations change its expression, by introducing random mutations into 8 yeast strains each carrying a single genetic variant responsible for altering the expression level of *TDH3* and quantifying both the mean expression level and expression noise for resulting mutagenized cells in each of the 8 genetic backgrounds. I found that the lab strain BY was less sensitive to random mutations on the mean expression level, compared to other genotypes carrying genetic variants. Also, I found that relationships between effects of random mutations on the mean level of expression and expression noise depend on the existing genetic variants. In addition, I found that the sensitivity to random mutations on mean level of expression was positively correlated with the expression noise for strains carrying genetic variants in the *TDH3* promoter. This study demonstrates that various aspects of how random mutations alter the expression of a single gene are modified by existing genetic changes that disrupt the transcriptional regulation.. Taken together, my thesis work demonstrates that the transcriptional regulatory network provides an informative context to study the evolution of gene expression, in the sense that both the process of the accumulation of genetic variations and formation of the ultimate evolutionary patterns are potentially affected by the interactions within the network.

Chapter I

Introduction

The evolution of gene expression contributes to phenotypic evolution

One of the most important questions in evolutionary biology is to understand the genetic basis for phenotypic evolution. Knowing what genetic changes are responsible for the observed phenotypic variations in natural populations could help researchers gain new insights into how evolution proceeds at the molecular level, and this information could improve our understanding on predictability of the evolution process. All genetic variations could be roughly classified into two categories: coding changes, which are mutations in protein coding sequences, and non-coding changes. Early studies in genetics and evolutionary biology have demonstrated that coding changes are important sources for the phenotypic variation. One of the most famous examples is the "*white*" gene in *Drosophila melanogaster* (Morgan 1910). Flies carrying mutations in the *white* gene have white eyes instead of normal red eyes. Due to the strong phenotypic consequences caused by coding changes, as well as genetic tools developed to examine existence of coding changes, evolutionary biologists had the strong belief that the genetic differences causing structural and functional variations in proteins served as important sources for the phenotypic evolution. However, since detailed descriptions on gene

expression were limited, it was not clear whether non-coding changes might also contribute to the phenotypic evolution.

After the 1950s, as soon as biologists started to unravel the underlying molecular mechanisms for the gene expression regulation, there were speculations on the possibility that genetic changes not in coding sequences could as well played important roles in the phenotypic evolution. For example, Jacob and Monod proposed the hypothesis that mutations in operators might be important in the evolution of prokaryotic gene expression (Monod and Jacob 1961). In the 1970s, two pioneer papers have been considered now to provide the most important conceptual basis for the importance of non-coding changes in phenotypic evolution. In 1969, based on the discovery that a large proportion of eukaryotic genomes are repetitive sequences, Britten and Davidson developed a simple gene regulation model (Britten and Davidson 1969). And from this model, they discussed that changes in gene expression regulation might play an important role in phenotypic evolution (Britten and Davidson 1971). In the second paper, King and Wilson found out that sequences of homologous proteins in humans and chimpanzees were almost identical, the fact of which suggested that humans and chimpanzees are not distinguished because of "human proteins" and "chimpanzees proteins" (King and Wilson 1975). The observed phenotypic differences between humans and chimpanzees could not be attributed to limited variations in coding sequences.

Although both studies brought up the implications that non-coding changes might be critical for the phenotypic evolution, they did not draw much attention due to the lack of the empirical evidence. In the recent decades, a large number of case studies have extensively shown that genetic changes that result in evolution of the regulation of gene

expression are important for phenotypic evolution (Hoekstra and Coyne 2007; Carroll 2008; Stern and Orgogozo 2008; Stern and Orgogozo 2009). I will use three representative examples to illustrate the tight connections between the evolution in gene expression and the phenotypic evolution.

In his *On the Origins of Species*, Darwin used the phenotypic differences of finches on Galapagos Island and Cocos Island to illustrate his theory of natural selection. One interesting observation was that the shapes and sizes of finch beak were different among populations on different islands. Darwin hypothesized that those differences were the consequences of adaptation to food types available on different islands. Abzhanov et al (2004) studied the developmental basis of the phenotypic differences among different finch populations and found out that expression level of the gene *Bone morphogenetic protein 4* (*BMP4*) correlates with beak length among different populations. They also showed that mis-expression *of BMP4* in chicken embryo could alter the beak morphology, suggesting that variations in expression level of *BMP4* might be the true cause of the phenotypic evolution of beak morphology in finch populations.

The pelvic apparatus in threespine sticklebacks is another widely used model to study the genetic basis for phenotypic evolution. Marine sticklebacks have developed a prominent pelvic skeleton, which could protect the fishes against soft-mouthed predators (Reimchen 1983). It has been found that multiple freshwater stickleback species partially or completely lost their pelvic structures (BELL 2008). By using whole genome linkage mapping approach, one chromosome region containing the gene *Pitx1* has been identified to explain over half of the variance in pelvic size among different stickleback populations (Cresko et al. 2004; Shapiro et al. 2004; Coyle et al. 2007). Compared to marine

stickleback populations, pelvic-reduced freshwater populations showed no coding changes in *Pitx1* gene (Shapiro et al. 2004). Instead, pelvic-reduced fishes partially or completely lost *Pitx1* expression in developing pelvic structures (Cole et al. 2003; Shapiro et al. 2004). Chan et al (Chan et al. 2010) identified regulatory mutations in a tissue specific enhancer of *Pitx1* gene in pelvic-reduced stickleback populations that are responsible for the changed expression pattern. They also found molecular signatures of positive selection on those mutations in freshwater stickleback populations. This story demonstrates that non-coding changes resulting in the evolution of gene expression are utilized as sources for phenotypic evolution in natural populations.

A recent study on limb development in vertebrates also illustrates the importance of regulatory mutations in morphological evolution (Kvon et al. 2016). The authors identified regulatory changes in the ZRS enhancer controlling the expression level of the gene *Shh* that are responsible for reduction or truncation of limb development in snake species. Those changes caused loss of expression of *Shh* during limb development, and in advance leaded to loss or reduction of leg structures in snake species. This study is another well-documented example suggesting that the evolution in gene expression level is a common mechanism for the evolution of morphological traits.

With more and more evidence suggesting that evolution of the regulation of gene expression are responsible for phenotypic evolution, there has been debate over whether coding changes or regulatory changes are more important for phenotypic evolution (Carroll 2008; Hoekstra and Coyne 2007). Although evidence for supporting regulatory changes as the major contributor to phenotypic evolution is accumulating quickly (Wray 2007; Carroll 2008; Stern and Orgogozo 2008; Wittkopp and Kalay 2011), this might be

due to the bias in techniques widely used to look for the genetic basis of phenotypic variation and differences in interpretation of the data (Hoekstra and Coyne 2007). Despite those potential caveats, the claim that regulatory changes play important roles in the phenotypic evolution demonstrates the necessity of understanding the regulation of gene expression in the goal of looking for the genetic basis of phenotypic evolution.

Interactions between transcription factors and *cis*-elements are important in the regulation of gene expression

The molecular mechanisms of the regulation of gene expression have been extensively studied over the last century. The process of gene expression is composed of several steps, including transcription, translation and other intermediate steps. Transcription, as the first step in gene expression, has been demonstrated to determine the temporal and spatial patterns of gene expression. To achieve the precise regulation of transcription, one of the commonly utilized mechanisms is that specific transcription factors bind to DNA sequences (*cis* elements) at the prescribed time and locations, and then recruit basal transcriptional machinery to the promoter region. Knowing what, when and where different transcription factors would regulate a gene is thus an important task for a better understanding of the regulation of transcription. One of the earliest studies on the regulation of the gene expression is the Lac operon system from Jacob and Monod (Jacob and Monod 1961). In this study, Jacob and Monod brought up the concept of "operator", which was a type of *cis*-elements in prokaryotes bound by transcription factors. Since then, an important part of transcriptional regulation studies is to look for transcription

factors regulating the genes under interest, as well as to figure out where and when the binding events occur.

Developmental biology is one of the fields that benefit the most from studies on the regulation of gene expression. In development, the gene expression is under precise control to achieve high precision in the formation of various morphological traits. In many developmental systems, the expression levels of multiple genes are regulated by the same sets of transcription factors. Focusing on the regulation of any single gene in this scenario could not provide sufficient information to understand how the final morphological characteristics are produced. One example to illustrate the complexity of the gene regulation in development is the early embryo development (Davidson et al. 2002; Erwin and Davidson 2009). Studies in multiple species demonstrate that setup of the body plan in the embryo stage is accomplished through precise temporal and spatial organizations of expression of many genes, and this coordinated transcriptional regulation is achieved through utilization of the same set of transcription factors for different genes. Even a complete description of the expression pattern of a single gene or transcription factor could not fully explain the characteristics of the final morphological traits. For the purpose of understanding developmental processes in a more systematic way, Davidson suggested the use of a "network" to represent interactions between transcription factors and target genes and claimed, "The architecture reveals features that can never be appreciated at any other level of analysis" (Levine and Davidson 2005). Since then, biological networks have been a very important research topic and have been shown to provide useful information in understanding the regulation of gene expression in many biological systems.

Transcriptional regulatory networks: basic concepts

A network is a collection of objects, which are called nodes, and descriptions of relationships among those nodes, which are called edges. Network is a very popular research topic in math, computer science, and many other social sciences. Using networks to represent interactions among objects has been shown to improve the understanding of the underlying system in many fields, including biology (Barabási and Oltvai 2004; Yu et al. 2013). The biological network used to study the regulation of gene expression is called Transcriptional Regulatory Network.. A transcriptional regulatory network represents regulatory interactions between transcription factors and their regulated genes. A node in a transcriptional regulatory network is a gene, either a transcription factor or a regulated gene. If a transcription factor directly regulates another gene through binding to a *cis* element, then an edge pointing from the transcription factor towards the regulated gene is present in the network. Each edge could be associated with a weight reflecting extra information, such as a numeric value indicating the strength of the binding, or a binary value indicating whether the transcriptional factor is an activator or a repressor.

Constructing a transcriptional regulatory network

Traditionally, the edges in a transcriptional regulatory network are discovered through detailed biochemical and molecular manipulation of the gene under interest and the candidate transcription factors. For example, Jacob and Monod discovered transcription factors for Lac operon through a series of genetic analyses, in which they

disrupted the function of transcription factors and checked their impact on the expression level of the Lac operon. The only disadvantage of this approach is that preexisting knowledge of potential regulators is required for the design of the experiments.

Thanks to the fast development of biotechnology, especially the second generation sequencing, biologists can now build transcriptional regulatory networks involving a large number of genes. Chromatin-Immunoprecipitation (ChIP), coupled with high-throughput techniques to find out binding locations (such as microarray and DNA-seq), provides a direct way to find out where a transcriptional factor could bind. Sequencing the whole transcriptome in response to a mutated version of a particular transcription factor provides a high-throughput replacement for the more traditional candidate gene perturbation approach, in which the impacts on the expression levels of other genes upon perturbing the function of the transcription factor under interest could be examined at the same time. Besides experimental approaches, development of bioinformatics methods has led to the *in silico* prediction of transcription factor binding sites in a DNA sequence. The rationale behind this approach is that the *cis* elements bound by a specific transcription factor have a preferential combination of base pairs (Schneider et al. 1986; Stormo and Hartzell 1989; Stormo 2000). By using randomly synthesized probes, estimates from ChIP experiments and other data sources, researchers could estimate the sequence preference of a particular transcriptional factor. The estimated preferred sequence is called binding "motif" (Stormo 2000). It is usually ~6-14bp long in most organisms, and in the position of each base pair is a weighted combination of all 4 possible nucleotides, with the weight representing the preference of each nucleotide at that position. There are many different motif databases generated by various labs. With

appropriate statistical models (Stormo 2000; Jayaram and Usvyat 2016), researchers can scan the genome to look for potential binding sites for a specific transcription factor by using the information stored in binding motifs.

Although all those high-throughput methods make it possible to build a whole-genome regulatory network, each of them has some drawbacks that could lead to imprecise predictions. For example, it is well-known that ChIP type experiments suffer from both high false-positive rate and false negative rate (Park 2009; Bailey et al. 2013). Specifically, binding peaks estimated from ChIP experiments could be non-functional, and this can depend on the strength of crosslinking used in the experiment. Also, weak direct binding events might not be captured by ChIP experiments. High-throughput functional assays, like RNA-seq in transcription factor knockout strains, provide evidence for direct regulatory interactions as well as indirect interactions. A more serious problem comes from the fact that it is not known whether the functional perturbation of a regulator would result in a detectable impact on the expression levels of its targets under the experimental conditions. It is suggested in multiple studies that the expression level of a gene could be robust to the perturbation in its regulators (Macneil and Walhout 2011; Steinacher et al. 2016). Finally, motif-based binding sites searching methods are limited in precision due to the inaccuracy in motif estimation (Simcha et al. 2012). In addition, it is not clear what standard one should use to assign the predicted binding sites to the target genes. The state-of-art approach is to assign the binding motifs to the closest gene, or to genes within a certain distance threshold. However, studies using new techniques such as Hi-C or 3D chromosome conformation showed that the *cis* elements located in enhancers could be far from their regulated genes (Anon 1993; Lieberman-Aiden et al. 2009; Rao et

al. 2015). Thus, the common practice of assigning motifs to the genes in close range might give misleading predictions. Recently, several groups used conservation in sequence estimated from comparisons across related species to validate the functional importance of the predicted motifs to improve the accuracy of the detection methods (Stark et al. 2007; Daily et al. 2011).

In the most recent decade, researchers proposed the idea of using "wisdom of crowds" to construct large scale transcriptional regulatory network (Marbach, Costello, et al. 2012). The basic rationale behind the "wisdom of crowds" is very simple: if an edge is predicted from different methods, then it is more likely that the regulatory interaction is present in the organism. One example is the construction of the *Drosophila melanogaster* regulatory network by using multiple sources of data generated from modENCODE (Marbach, Roy, et al. 2012). In this study, the authors combined four sources of features: the physical binding datasets (ChIP experiments), the conserved motif instances estimated from aligning 12 closely related *Drosophila* species from DGRP project (Stark et al. 2007), the correlation of the expression of all genes across multiple developmental stages, and the correlation of the histone modification markers from multiple regions of gene body. By using a simple statistical learning method, the resulted transcriptional regulatory network had a relatively good performance in multiple validation metrics. Also, the authors utilized the network to predict the functions of genes with no annotation available as well as predict expression level of target genes based on predicted regulators. They found out that the outcomes were better than regulatory networks imputed from a single source of data, which suggested that the

principle of "wisdom of crowds" provided a useful conceptual basis to construct regulatory networks.

The so-called whole genome transcriptional regulatory networks constructed from those high-throughput methods described above ignore developmental stages, environment conditions and tissue specificity. Although suffering from different levels of imprecision, utilizing the regulatory network could help us gain new understandings of the regulation and evolution of gene expression, which is the next topic.

Structure properties of transcriptional regulatory networks have functional impacts on gene expression regulation and evolution

Studies in developmental biology, systems biology and evolutionary biology in recent decades have all shown that using biological networks provides new insights in understanding the underlying biological systems. For the transcriptional regulatory network, the major focus is on whether and how specific patterns of the structural organizations among nodes bring us knowledge on observed patterns related to the regulation of gene expression..

It has been reported in different systems that the transcriptional regulatory network has a hierarchical structure (Gerstein et al. 2012; Jothi et al. 2009; Cosentino Lagomarsino et al. 2007), in which multiple layers of regulatory interactions are present in the network. Some transcription factors are near the top of the hierarchy, and they are called "master regulators", which control the expression level of other genes with different functions. Transcription factors in the intermediate layer in the hierarchy,

together with their targets, produce different "modules" that execute relatively a smaller number of functions (Erwin and Davidson 2009). It is suggested that modularity in a transcriptional regulatory network could benefit coordination of different biological process during metabolism, cellular signaling, and development (Davidson et al. 2002; Barabási and Oltvai 2004; Alon 2007).

Besides recurrent patterns in the large-scale organizations, researchers also recognized small "prototype" circuits, which are called "network motifs" (Alon 2007). One example is the so-called "feed forward loop" (FFL). This motif is composed of 3 genes, including two transcription factors A and B, and a target gene C. In one type of FFL motifs, A and B both activate C, while A also activates B. By organizing regulatory interactions in this way, the expression level of gene C can response to the changes in the expression level of the gene A in a pulse manner (Alon 2007). Many studies have shown that different network motifs produce different dynamics of the gene expression level (Shen-Orr et al. 2002; Eichenberger et al. 2004; Odom et al. 2004; Alon 2007). Furthermore, the resulted dynamics are widely used in many different biological processes. For example, the motif called "bi-stable switch" is used in different systems of cell fate determination, where precise boundaries of gene expression are required (Xiong and Ferrell 2003; Brandman et al. 2005; Canela-Xandri et al. 2008; Andrecut et al. 2011).

The structural properties of the regulatory network not only influence the dynamics of the regulation of gene expression, but might also affect the evolution of gene expression. Due to the accumulation of mutations in both *cis* elements and transcription factors, the structure of a transcriptional regulatory network is evolving all the time.

From network science, the future state of each node (in this case, the expression level of genes in the network) depends on what partners it interacts with and how those interactions are organized (Newman 2013). The question of understanding the evolution of gene expression under the context of the transcriptional regulatory network is thus to ask the question that whether the structural organizations of the network could influence the evolutionary changes of the expression levels of the genes within the network.

One of the interesting arguments about the relationship between the evolution of gene expression and the regulatory network comes from the discussion on the predictability of genetic evolution (Stern and Orgogozo 2009). The rationale behind the hypothesis that the genetic basis of phenotypic evolution is predictable stems from the observations that mutations responsible for phenotypic evolution are not randomly distributed (Stern and Orgogozo 2009; Carroll 2008). What's more, multiple case studies show that even though there are many mutations that could achieve the phenotypic divergence observed in nature, only a few of them have been actually used by various species during parallel evolution on the same phenotype (Levy and Dean 1998; Shindo et al. 2005; McGregor et al. 2007). Stern and Orgogozo (2009) used the *shavenbaby* story (McGregor et al. 2007) to illustrate the idea that the position of a transcription factor in the transcriptional regulatory network could affect how its expression would evolve. The *shavenbaby* gene contributes to trichome structure in *Drosophila* species. It was observed that changes in *cis*-elements upstream of *shavenbaby* were responsible for trichome divergence in multiple *Drosophila* species. Although similar phenotypic variations might be achieved from genetic changes in other genes, there would be other side effects produced at the same time. Genetic changes in genes upstream of *svb* in the

network could cause complete disruption of the development of many other organs. However, genetic changes in genes downstream of *svb* in the network are not sufficient to change the complete trichome structure. Thus, *shavenbaby* is expected to be a good target if variations in trichome structures in *Drosophila* species are required in different environments.

In the network science, a node that is critical for information flow is called a network "hub". Network hubs receive input from the upstream nodes and pass the information to the downstream nodes. The gene *shavenbaby* could be considered as a network hub in trichome development. If natural selection favors changes in trichome structure, *svb* might be expected to be the hotspot for regulatory evolution. However, if selection provides constraint so that changing the underlying phenotypes is lethal, then the network hub might have a stabilized expression level. He et al (He and Zhang 2006) found that network hubs in the protein-protein interaction (PPI) network with large number of interacting partners, tend to be essential genes in yeast. Although this is not an example in a transcription regulatory network, this study still emphasizes the idea that the network context of a gene could affect its evolution.

In the Stern and Orgogozo paper, the authors also discussed characteristics that could affect predictions of genetic changes underlying phenotypic evolution (Stern and Orgogozo 2009). One of them is pleiotropy. Pleiotropy refers to the phenomenon that changes in the function or the expression level of a gene result in changes in multiple phenotypes. It is hypothesized that the expression levels of genes with higher pleiotropic effect are more stable than those with lower pleiotropic effect, because changes in the expression level of the former group of genes could impact more traits, which in turn are

more likely to bring detrimental effects to the organisms (Paaby and Rockman 2013). The number of targets of a transcription factor within the transcriptional regulatory network could be considered as an approximation of the level of pleiotropy, especially if researchers could infer how many different phenotypes are affected based on the phenotypic functions of its targets. By combining the regulatory network and divergence in the expression of genes across related species, one can directly test the predictions from theory of pleiotropy.

Not only could the number of targets affects the evolution of the expression, the number of regulators of a gene within a transcriptional regulatory network might also influence the evolution of the expression level. Several studies have shown that master regulators and core regulators in development have more regulators in the transcriptional regulatory network (Borneman et al. 2006; Vermeirssen et al. 2007), and those genes tend to have less variation in gene expression (Batada and Hurst 2007). These observations imply that gene expression might be more stable for genes with more regulators. However, compared to genes with lower number of regulators,, a random mutation has higher chance to hit regulators of genes with higher number of incoming connections, which suggests that the expression level should be more diverged for genes with more regulators. Landry et al (Landry et al. 2007) showed that sensitivity of the expression level towards random mutations is positively correlated with number of regulators in a mutation accumulation study in *S. cerevisiae*. It is thus interesting to test whether genes with more regulators in a regulatory network are more or less likely to change expression over evolutionary time, and whether the conclusions in different species groups are consistent or not.

Evolution of transcriptional regulatory networks

In the recent decades, the evolution of the structure of the transcriptional regulatory network has become a popular research topic in evolutionary biology. Since it is difficult to obtain information of the structure of the regulatory network in diverged species due to the lack of available data, most experimental studies on this topic focus on providing complete descriptions of changes in small regulatory networks responsible for specific diverged phenotypes. Researchers use comparative studies on the evolution of the regulatory network to understand the general strategies used by organisms to expand or reorganize the regulatory interactions. Another angle to examine the evolution of the regulatory network is to use mathematical modeling as well as *in silico* simulation to explore how the structural properties commonly observed for biological networks are generated in evolution.

Using comparative studies to examine the evolution of the regulatory network generates many insights on strategies utilized in changing the structure of the network. It has been demonstrated that that network circuits responsible for core developmental processes are relatively conserved across evolutionary time (Rebeiz et al. 2015; Thompson et al. 2015). For example, regulatory networks responsible for endomesoderm development in sea urchin and sea star, which diverged 400 million years ago, have exactly the same components and structural organizations (Hinman et al. 2003; McCauley et al. 2010). Even in the scenario that phenotypic innovation is preferred by natural selection, it was found that instead of generating all new interactions among unrelated genes from scratch, in many cases phenotypic innovation is achieved through reusing existing regulatory network circuits with only a few new connections generated

to depict the timing and placing of the activation of the old circuits in new context (Thompson et al. 2015; Rebeiz et al. 2015), and this process is called co-option. One of the most well-known examples is the horn development in beetles, in which a small network circuit, including the transcription factor *Distal-less*, is reused in multiple segments in beetles to generate a varied number of appendages across different species (Moczek et al. 2006). All those examples suggest that the evolution of the network structures follows a modular manner, in which expansion and reorganization of the structure of the regulatory network are achieved through using existing regulatory modules while only adding new regulatory interactions that put the old modules in appropriate temporal and spatial context when necessary..

However, it is still not clear how the existing modules form in evolution. What's more, biological networks in different systems share many interesting properties. As discussed above, the transcriptional regulatory network shows specific structural properties, including modularity, hierarchy, and repeated occurrence of network motifs. A natural question to ask is whether those properties are the results of natural selection or created by neutral processes. Several researchers believed that natural selection is important in the formation of those properties in the regulatory network (Davidson et al. 2002; Barabási and Oltvai 2004; Alon 2007; Rebeiz et al. 2015). However, using simple mathematical models based on random gain and loss of regulatory interactions, Lynch (2007) showed that the topological properties of the transcriptional regulatory network could be generated only through mutations and duplications, without the need of natural selection. Interestingly, it is found that large-scale network reorganization is often caused by gene duplication followed by random gain of binding sites for duplicated

transcriptional factors or in the *cis* regulatory regions of duplicated genes (Voordeckers et al. 2015). Although it is not clear whether selection or drift are more important in the evolution of the network structure, results from extensive comparative studies and theoretical analyses could generate hypothesis that can be tested when more datasets are available for more diverged species.

From the above discussion, the transcriptional regulatory network plays an important role in both the regulation and the evolution of gene expression. However, the regulatory network only imposes constraints on the evolution of gene expression. It is not the driving force for the evolutionary process. The evolutionary fate of the expression is determined by both the existing variations within the population and the selection constraints imposed from the environment as well as other evolutionary forces. Thus, a complete understanding of the evolution of gene expression requires a better understanding of both the mutational process that generates the variations and the evolutionary forces that act on the variations.

Understanding the phenotypic effect of new mutations improves our understanding of phenotypic evolution

Although much more precise from the mechanistic perspective, our current understanding of the phenotypic evolution is similar to Darwin's conceptual framework (Darwin 1859). First, phenotypic variation accumulates within a population. Then evolutionary forces, such as selection or drift, would then act on those phenotypic variations. The final outcome is the collection of phenotypic variations observed in nature. Mutations

determine what phenotypes are available for evolutionary forces, and these evolutionary forces determine what phenotypes will be retained over time.

Since natural selection and genetic drift constantly remove phenotypic variation out of the population, mutations are thought to be critical for phenotypic evolution in the sense that without newly generated mutations, evolution will halt. However, historically, attitudes towards the importance of mutations in evolution have experienced drastic changes. In Darwin's era, the majority of the biologists had imprecise understandings about how genetic variations were generated, because the basic principles of genetics and molecular biology were not accessible to them (Nei 2013). This condition did not change until the rediscovery of Mendel's law, and efforts from early quantitative evolutionary biologists showed that Mendel's inheritance law is not only applicable to discontinuous traits, but also correct for continuous traits. The combination of the two shaped the primary form of the two steps process mentioned above: discrete genetic variations first occur in inheritance units, and then natural selection or genetic drift act on those variations.

However, the role of new mutations in evolution was not considered important before the 1960s. There are two major reasons (Nei 2013). First, early quantitative genetic models showed that natural selection and genetic drift had much stronger power in determining allele frequency changes in a population, while differences in frequency of starting mutations only played a minor role. Second, it was widely accepted that mutations were abundant in a population, so that natural selection and genetic drift have access to genetic variations with any possible phenotypic effects. Based on the above arguments, natural selection and genetic drift have long been thought of as the major

forces of phenotypic evolution, while the mutational process was only considered as providing input for evolutionary forces to act upon (Nei 2013).

In 1962, Zuckerlandl and Pauling found out that the rate of evolution of human hemoglobins correlated with the mutation rate of coding sequences (Zuckerkandl and Pauling 1962). This conclusion was not consistent with the dominant view of natural selection as the primary force of evolution. If that is true, the rate of evolution should correlate with the rate of environment changes instead of mutation rate, which is an important characteristic of the mutational process (Nei 2013). However, this study did not provide direct evidence to support the importance of mutational process in evolution.

Part of the reason why the mutational process was thought to be of little importance in evolution was the lack of knowledge on the underlying molecular mechanisms. With the recent advances in the knowledge of genetic basis of many biological processes, researchers are gradually realizing that a complete description of the mutational process is also necessary for deciphering or predicting the phenotypic evolution. For example, natural selection alone cannot explain the fact that *svb* was used independently by multiple *Drosophila* species for phenotypic innovation on trichome development (Shapiro et al. 2004). Natural selection, which are constraints imposed on phenotypes of the organisms by the environment, is blind to the underlying genetic basis. Following this logic, it is recognized from studies of developmental biology that developmental processes are constrained by their molecular process (Vrba and Eldredge 1984), which suggests that the specific effect and identity of a mutation is also important for understanding its potential for phenotypic evolution (Hall 2003).

Also, recent studies showed that previous understanding of the phenotypic effects of new mutations from quantitative genetics might not be consistent with empirical observations. It was found that predicted properties of the distribution of fitness effect of new mutations (Eyre-Walker and Keightley 2007; Rice et al. 2015) are not consistent with empirical data (Eyre-Walker et al. 2006; Rokyta et al. 2008; Levy et al. 2015). The inconsistency between theoretical predictions and experimental observations highlight the necessity of determining phenotypic effects of random mutations experimentally.

Genetic background could influence effect of new mutations

The underlying genetic background affects phenotypic effects of new mutations. More specifically, the existence of other genetic changes within the genome could change the phenotypic effects of new mutations, a phenomenon known as epistasis. Epistasis is widely present among mutations, and this is illustrated by the studies showing that different genetic backgrounds can modulate the effects of new mutations (e.g., (McKenzie et al. 1982; Remold and Lenski 2004; Milloz et al. 2008; Dworkin et al. 2009; Wang et al. 2013). Also, genetic changes that influence effects of other mutations might segregate in natural populations. The influence of those genetic changes might not be obvious in some environments, but their impacts on new mutations are illustrated once new experimental conditions are introduced. Those genetic changes are called cryptic genetic variations and are described in multiple studies (Gibson and Dworkin 2004; Duveau and Félix 2012; Ledón-Rettig and Pfennig 2014). Overall, all those studies

emphasize the importance of the underlying genetic background when researchers are studying effects of new mutations.

One important discovery from the recent experimental evolution studies is that pre-existing mutations have a huge impact on the evolutionary fate for new mutations. For example, Weinreich et al (Weinreich et al. 2006) identified five point mutations in β-lactamase that together can increase the bacterial resistance to antibiotics. However, by reconstructing all 120 possible orders of occurrences of those five mutations, only 18 paths were permitted when applying the selective pressure using antibiotics, and one of the five mutations always appeared first in all 18 paths. Another interesting study on hormone-receptor co-evolution in vertebrates illustrated that evolutionary transitions of receptor specificity towards ligands were dependent on the random occurrence of several key amino acid changes within the protein (Bridgham et al. 2006). The above two case studies both suggest that the trajectory of how a biological system evolves in the molecular level not only depends on the constraints from the environment, but also depends on what variations are already present in the genome. This concept is supported by multiple recent experimental evolution studies in unisexual organisms (reviewed in (Lang and Desai 2014)). Discoveries from all those studies suggest that effect sizes, fitness effect, and identities of new mutations that will arise in short-term evolution largely depend on pre-existing genetic changes present in the genome.

From the above discussion, it is important to understand not only phenotypic effects of new mutations, but also to collect this information in multiple genetic backgrounds.

Thesis Outline

In this thesis, I examine whether and how transcriptional regulatory networks can impact both observed gene expression evolution and underlying mutational process.

In chapter 2, I combined a transcriptional regulatory network constructed in *D. melanogaster* with gene expression evolution data from three within/between species comparisons among multiple species in *Drosophila* group (*D. melanogaster, D. simulans, D. sechellia*), in order to examine whether connectivity properties within regulatory network might influence observed gene expression evolution pattern. I showed that increasing number of regulators (in-degree) for a gene was associated with decreasing expression divergence over time. This observation suggests that the high number of regulators could provide robustness to gene expression variation over evolutionary time. Also, based on prediction from the theory of pleiotropy that transcription factors affecting multiple traits have restricted chance of divergence on gene expression level, I checked whether the number of target genes (out-degree) for a transcription factor was associated with gene expression evolution and found no statistical evidence to support the prediction.

In chapter 3, I used datasets from *Saccharomyces* species to examine whether properties of connectivity in a transcriptional regulatory network have a consistent relationship with gene expression in different groups of related species. I first reconstructed a transcriptional regulatory network for *S. cerevisiae* by following methods developed in Marbach et al (Marbach, Roy, et al. 2012), incorporating multiple sources of datasets

representing different aspects of regulatory interactions. I showed that the reconstructed regulatory network captured more functional informative regulatory interactions than previous regulatory networks. I then combined the reconstructed regulatory network with gene expression evolution data from four comparisons within and between species in the *Saccharomyces* group (*S. cerevisiae, S. paradoxus, S. mikatae* and *S. bayanus*). I showed that the increasing in-degree was associated with both increasing probability and magnitude of differences in gene expression / *cis* regulation between two strains of *S. cerevisiae,* while no significant relationships were detected in comparisons between diverged species. This conclusion was inconsistent with findings in *Drosophila* species. By comparing the two studies and looking for possible explanations for the observed differences, I argued that organization of network in different organisms might have different properties, which could impact gene expression evolution over evolutionary time.

In chapter 4, I used a fluorescent reporter gene to specifically study how genetic background / pre-existing genetic variants could affect mutational effects on both mean level of expression and expression noise. I found that the mean level of expression of the yeast lab strain was more robust to random mutations than other genotypes carrying existing genetic variants disrupting expression of the reporter gene, while this was not true for expression noise. In addition, I found that the relationships between the mean level of gene expression and expression noise were different among different genetic backgrounds, suggesting that prior genetic variants could impact combinatory effects of new mutations on mean level of expression and expression noise. Finally, I showed that

increasing variation on mutational effects for the mean level of expression was positively associated with increasing expression noise at least for starting genetic variants in *cis* elements, suggesting that sensitivity to random mutations of gene expression level might be correlated with sensitivity to internal molecular fluctuation during the transcription process.

Refenrences

Abzhanov A, Protas M, Grant BR, Grant PR, Tabin CJ. 2004. Bmp4 and Morphological Variation of Beaks in Darwin's Finches. Science 305:1462–1465.

Alon U. 2007. Network motifs: theory and experimental approaches. Nature Reviews Genetics 8:450–461.

Andrecut M, Halley JD, Winkler DA, Huang S. 2011. A General Model for Binary Cell Fate Decision Gene Circuits with Degeneracy: Indeterminacy and Switch Behavior in the Absence of Cooperativity.Monk N, editor. PLoS ONE 6:e19358.

Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, Madrigal P, Taslim C, Zhang J. 2013. Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data.Lewitter F, editor. PLoS Comput Biol 9:e1003326.

Barabási A-L, Oltvai ZN. 2004. Network biology: understanding the cell's functional organization. Nature Reviews Genetics 5:101–113.

Batada NN, Hurst LD. 2007. Evolution of chromosome organization driven by selection for reduced gene expression noise. Nature Genetics 39:945–949.

BELL MA. 2008. Interacting evolutionary constraints in pelvic reduction of threespine sticklebacks, Gasterosteus aculeatus (Pisces, Gasterosteidae). Biological Journal of the Linnean Society 31:347–382.

Borneman AR, Leigh-Bell JA, Yu H, Bertone P, Gerstein M, Snyder M. 2006. Target hub proteins serve as master regulators of development in yeast. Genes Dev. 20:435–448.

Brandman O, Ferrell JE, Li R, Meyer T. 2005. Interlinked Fast and Slow Positive Feedback Loops Drive Reliable Cell Decisions. Science 310:496–498.

Bridgham JT, Carroll SM, Thornton JW. 2006. Evolution of Hormone-Receptor Complexity by Molecular Exploitation. Science 312:97–101.

Britten RJ, Davidson EH. 1969. Gene regulation for higher cells: a theory. Science.

Britten RJ, Davidson EH. 1971. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. The Quarterly Review of Biology 46:111–138.

Canela-Xandri O, Sagués F, Reigada R, Buceta J. 2008. A Spatial Toggle Switch Drives Boundary Formation in Development. Biophysical Journal 95:5111–5120.

Carroll SB. 2008. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. Cell 134:25–36.

Chan YF, Marks ME, Jones FC, Villarreal G, Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J, et al. 2010. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science 327:302–305.

Cole NJ, Tanaka M, Prescott A, Tickle C. 2003. Expression of limb initiation genes and clues to the morphological diversification of threespine stickleback. Current Biology 13:R951–R952.

Cosentino Lagomarsino M, Jona P, Bassetti B, Isambert H. 2007. Hierarchy and feedback in the evolution of the Escherichia coli transcription network. Proceedings of the National Academy of Sciences 104:5516–5520.

Coyle SM, Huntingford FA, Peichel CL. 2007. Parallel Evolution of Pitx1 Underlies Pelvic Reduction in Scottish Threespine Stickleback (Gasterosteus aculeatus). J Hered 98:581–586.

Cresko WA, Amores A, Wilson C. 2004. Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations.

Daily K, Patel VR, Rigor P, Xie X, Baldi P. 2011. MotifMap: integrative genome-wide maps of regulatory motif sites for model species. BMC Bioinformatics 12:495.

Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh C-H, Minokawa T, Amore G, Hinman V, Arenas-Mena C, et al. 2002. A Genomic Regulatory Network for Development. Science 295:1669–1678.

Duveau F, Félix M-A. 2012. Role of Pleiotropy in the Evolution of a Cryptic Developmental Variation in Caenorhabditis elegans.Noor MAF, editor. PLOS Biol 10:e1001230.

Dworkin I, Kennerly E, Tack D, Hutchinson J, Brown J, Mahaffey J, Gibson G. 2009. Genomic Consequences of Background Effects on scalloped Mutant Expressivity in

the Wing of Drosophila melanogaster. Genetics 181:1065–1076.

Eichenberger P, Fujita M, Jensen ST, Conlon EM, Rudner DZ, Wang ST, Ferguson C, Haga K, Sato T, Liu JS, et al. 2004. The Program of Gene Transcription for a Single Differentiating Cell Type during Sporulation in Bacillus subtilis.Jonathan A Eisen, editor. PLOS Biol 2:e328.

Erwin DH, Davidson EH. 2009. The evolution of hierarchical gene regulatory networks. Nature Reviews Genetics 10:141–148.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. Nature Reviews Genetics 8:610–618.

Eyre-Walker A, Woolfit M, Phelps T. 2006. The Distribution of Fitness Effects of New Deleterious Amino Acid Mutations in Humans. Genetics 173:891–900.

Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan K-K, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, et al. 2012. Architecture of the human regulatory network derived from ENCODE data. Nature 489:91–100.

Gibson G, Dworkin I. 2004. Uncovering cryptic genetic variation. Nature Reviews Genetics 5:681–690.

Hall BK. 2003. Evo-Devo: evolutionary developmental mechanisms. Int. J. Dev. Biol. 47:491–495.

He X, Zhang J. 2006. Why Do Hubs Tend to Be Essential in Protein Networks? PLoS Genet. 2:e88.

Hinman VF, Nguyen AT, Cameron RA, Davidson EH. 2003. Developmental gene regulatory network architecture across 500 million years of echinoderm evolution. Proceedings of the National Academy of Sciences 100:13356–13361.

Hoekstra HE, Coyne JA. 2007. THE LOCUS OF EVOLUTION: EVO DEVO AND THE GENETICS OF ADAPTATION. Evolution 61:995–1016.

Jacob F, Monod J. 1961. On the Regulation of Gene Activity. Cold Spring Harbor Symposia on Quantitative Biology 26:193–211.

Jayaram N, Usvyat D. 2016. Evaluating tools for transcription factor binding site prediction. BMC ….

Jothi R, Balaji S, Wuster A, Grochow JA, Gsponer J, Przytycka TM, Aravind L, Babu MM. 2009. Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture. Molecular Systems Biology 5:294.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees.

Kvon EZ, Kamneva OK, Melo US, Barozzi I, Osterwalder M, Mannion BJ, Tissières V, Pickle CS, Plajzer-Frick I, Lee EA, et al. 2016. Progressive Loss of Function in a Limb Enhancer during Snake Evolution. Cell 167:633–642.e11.

Landry CR, Lemos B, Rifkin SA, Dickinson WJ, Hartl DL. 2007. Genetic Properties Influencing the Evolvability of Gene Expression. Science 317:118–121.

Lang GI, Desai MM. 2014. The spectrum of adaptive mutations in experimental evolution. Genomics 104:412–416.

Ledón-Rettig CC, Pfennig DW. 2014. Cryptic genetic variation in natural populations: a predictive framework. Integrative and ….

Levine M, Davidson EH. 2005. Gene regulatory networks for development. Proceedings of the National Academy of Sciences 102:4936–4942.

Levy SF, Blundell JR, Venkataram S, Petrov DA, Fisher DS, Sherlock G. 2015. Quantitative evolutionary dynamics using high-resolution lineage tracking. Nature 519:181–186.

Levy Y, Dean C. 1998. The transition to flowering. Plant Cell 10:1973–1990.

Lieberman-Aiden E, Van Berkum NL, Williams L. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. ….

Lynch M. 2007. The evolution of genetic networks by non-adaptive processes. Nature Reviews Genetics 8:803–813.

Macneil LT, Walhout AJM. 2011. Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. Genome Res. 21:645–657.

Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Consortium TD, Kellis M, Collins JJ, et al. 2012. Wisdom of crowds for robust gene network inference. Nature Methods 9:796–804.

Marbach D, Roy S, Ay F, Meyer PE, Candeias R, Kahveci T, Bristow CA, Kellis M. 2012. Predictive regulatory models in Drosophila melanogaster by integrative inference of transcriptional networks. Genome Res. 22:1334–1349.

McCauley BS, Weideman EP, Hinman VF. 2010. A conserved gene regulatory network subcircuit drives different developmental fates in the vegetal pole of highly divergent echinoderm embryos. Developmental Biology 340:200–208.

McGregor AP, Orgogozo V, Delon I, Zanet J, Srinivasan DG, Payre F, Stern DL. 2007. Morphological evolution through multiple cis-regulatory mutations at a single gene. Nature 448:587–590.

McKenzie JA, Whitten MJ, Adena MA. 1982. The effect of genetic background on the

fitness of diazinon resistance genotypes of the Australian sheep blowfly, Lucilia cuprina. Heredity 49:1–9.

Milloz J, Duveau F, Nuez I, Félix M-A. 2008. Intraspecific evolution of the intercellular signaling network underlying a robust developmental system. Genes Dev. 22:3064–3075.

Moczek AP, Rose D, Sewell W, Kesselring BR. 2006. Conservation, innovation, and the evolution of horned beetle diversity. Dev Genes Evol 216:655–665.

Monod J, Jacob F. 1961. General Conclusions: Teleonomic Mechanisms in Cellular Metabolism, Growth, and Differentiation. Cold Spring Harbor Symposia on Quantitative Biology 26:389–401.

Morgan TH. 1910. SEX LIMITED INHERITANCE IN DROSOPHILA. Science 32:120–122.

Odom DT, Zizlsperger N, Gordon DB, Bell GW, Rinaldi NJ, Murray HL, Volkert TL, Schreiber J, Rolfe PA, Gifford DK, et al. 2004. Control of Pancreas and Liver Gene Expression by HNF Transcription Factors. Science 303:1378–1381.

Paaby AB, Rockman MV. 2013. The many faces of pleiotropy. Trends in Genetics 29:66–73.

Park PJ. 2009. ChIP|[ndash]|seq: advantages and challenges of a maturing technology. Nature Reviews Genetics 10:669–680.

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. 2015. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell 162:687–688.

Rebeiz M, Patel NH, Hinman VF. 2015. Unraveling the Tangled Skein: The Evolution of Transcriptional Regulatory Networks in Development. http://dx.doi.org/10.1146/annurev-genom-091212-153423 16:103–131.

Reimchen TE. 1983. Structural Relationships Between Spines and Lateral Plates in Threespine Stickleback (Gasterosteus aculeatus). Evolution 37:931.

Remold SK, Lenski RE. 2004. Pervasive joint influence of epistasis and plasticity on mutational effects in Escherichia coli. Nature Genetics 36:423–426.

Rice DP, Good BH, Desai MM. 2015. The Evolutionarily Stable Distribution of Fitness Effects. Genetics 200:321–329.

Rokyta DR, Beisel CJ, Joyce P, Ferris MT, Burch CL, Wichman HA. 2008. Beneficial Fitness Effects Are Not Exponential for Two Viruses. J Mol Evol 67:368–376.

Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. 1986. Information content of binding

sites on nucleotide sequences. J. Mol. Biol. 188:415–431.

Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, Jónsson B, Schluter D, Kingsley DM. 2004. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature 428:717–723.

Shen-Orr SS, Milo R, Mangan S, Alon U. 2002. Network motifs in the transcriptional regulation network of Escherichia coli. Nature Genetics 31:64–68.

Shindo C, Aranzana MJ, Lister C, Baxter C, Nicholls C, Nordborg M, Dean C. 2005. Role of FRIGIDA and FLOWERING LOCUS C in determining variation in flowering time of Arabidopsis. Plant Physiol. 138:1163–1173.

Simcha D, Price ND, Geman D. 2012. The Limits of De Novo DNA Motif Discovery.Peddada SD, editor. PLoS ONE 7:e47836.

Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN, et al. 2007. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature 450:219–232.

Steinacher A, Bates DG, Akman OE, Soyer OS. 2016. Nonlinear Dynamics in Gene Regulation Promote Robustness and Evolvability of Gene Expression Levels.Proulx SR, editor. PLoS ONE 11:e0153295.

Stern DL, Orgogozo V. 2008. THE LOCI OF EVOLUTION: HOW PREDICTABLE IS GENETIC EVOLUTION? Evolution 62:2155–2177.

Stern DL, Orgogozo V. 2009. Is Genetic Evolution Predictable? Science 323:746–751.

Stormo GD, Hartzell GW. 1989. Identifying protein-binding sites from unaligned DNA fragments. Proceedings of the National Academy of Sciences 86:1183–1187.

Stormo GD. 2000. DNA binding sites: representation and discovery. Bioinformatics 16:16–23.

Thompson D, Regev A, Roy S. 2015. Comparative Analysis of Gene Regulatory Networks: From Network Reconstruction to Evolution. Annual Review of Cell and Developmental Biology 31:399–428.

Vermeirssen V, Barrasa MI, Hidalgo CA, Babon JAB, Sequerra R, Doucette-Stamm L, Barabási A-L, Walhout AJM. 2007. Transcription factor modularity in a gene-centered C. elegans core neuronal protein-DNA interaction network. Genome Res. 17:1061–1071.

Voordeckers K, Pougach K, Verstrepen KJ. 2015. How do regulatory networks evolve and expand throughout evolution? Curr. Opin. Biotechnol. 34:180–188.

Vrba ES, Eldredge N. 1984. Individuals, hierarchies and processes: towards a more

complete evolutionary theory. Paleobiology 10:146–171.

Wang Y, Arenas CD, Stoebel DM, Cooper TF. 2013. Genetic background affects epistatic interactions between two beneficial mutations. Biology Letters 9:20120328–20120588.

Weinreich DM, Delaney NF, DePristo MA, Hartl DL. 2006. Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. Science 312:111–114.

Wittkopp PJ, Kalay G. 2011. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nature Reviews Genetics.

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. Nature Reviews Genetics 8:206–216.

Xiong W, Ferrell JE. 2003. A positive-feedback-based bistable |[lsquo]|memory module|[rsquo]| that governs a cell fate decision. Nature 426:460–465.

Yu D, Kim M, Xiao G, Hwang TH. 2013. Review of Biological Network Data and Its Applications. Genomics & Informatics 11:200–210.

Zuckerkandl E, Pauling L. 1962. Molecular disease, evolution and genetic heterogeneity.

1993. Interaction between transcription regulatory regions of prolactin chromatin.

Chapter II

The structure of the transcriptional regulatory network correlates with regulatory divergence in drosophila

Abstract

Transcriptional control of gene expression is regulated by biochemical interactions between *cis*-regulatory DNA sequences and *trans*-acting factors that form complex regulatory networks. Genetic changes affecting both *cis*- and *trans*-acting sequences in these networks have been shown to alter patterns of gene expression as well as higher-order organismal phenotypes. Here, we investigate how the structure of these regulatory networks relates to patterns of polymorphism and divergence in gene expression. To do this, we compared a transcriptional regulatory network inferred for *Drosophila melanogaster* to differences in gene regulation observed between two strains of *D. melanogaster* as well as between two pairs of closely related species: *Drosophila sechellia* and *Drosophila simulans,* and *D. simulans* and *D. melanogaster*. We found that the number of transcription factors predicted to directly regulate a gene ("in-degree") was negatively correlated with divergence in both gene expression (mRNA abundance) and *cis*-regulation. This observation suggests that the number of transcription factors directly regulating a gene's expression affects the conservation of *cis*-regulation and gene expression over evolutionary time. We also tested the hypothesis that transcription factors regulating more target genes (higher "out-degree") are less likely to evolve changes in

their *cis*-regulation and expression (presumably due to increased pleiotropy), but found little support for this predicted relationship. Taken together, these data show how the architecture of regulatory networks can influence regulatory evolution.

Introduction

Genetic changes that alter gene expression contribute to phenotypic evolution, thus understanding how gene expression is regulated and changes over evolutionary time is important for understanding how phenotypes evolve (Wray 2007; Carroll 2008; Stern and Orgogozo 2009; Wittkopp and Kalay 2011). The first step in gene expression is transcription, which is controlled by interactions between *trans*-acting transcription factors and *cis*-acting DNA sequences. Transcriptional regulatory networks summarize the connections between transcription factors and the genes that they regulate, known as their 'target genes' (Zhu et al. 2007). Because evolutionary changes arise within the context of these regulatory networks, the architecture of a regulatory network might make some types of changes more likely to evolve than others. Indeed, the connectivity of genes in a transcriptional regulatory network (i.e., the number of genes a gene regulates or is regulated by) has been found to correlate with evolutionary properties such as the rate of coding sequence evolution and gene duplication (e.g., Evangelisti and Wagner 2004; Jovelin and Phillips 2009).

Connectivity within a transcriptional regulatory network might also influence the evolution of transcriptional regulation itself. For example, the number of transcription factors regulating expression of a gene, a quantity known as 'in-degree', has been shown to positively correlate with plasticity in gene expression among environments (Promislow

2005) as well as mutational variance (Landry et al. 2007). This latter study, which examined the effects of new mutations arising in the near absence of selection on mRNA abundance, found that genes whose *cis*-regulatory elements had binding sites for more transcription factors were more likely to have their expression altered by new mutations, presumably because such genes had a larger mutational target size. It has also been suggested, however, that new mutations are less likely to alter expression of genes with many transcriptional regulators (higher in-degree) than genes with fewer transcriptional regulators (lower in-degree) because of robustness conferred by transcription factor binding sites with redundant or overlapping functions (Macneil and Walhout 2011). Natural selection is also expected to enforce greater constraints on expression of genes with higher in-degree because they tend to be key players in developmental pathways and changes in their expression tend to have large phenotypic consequences (Borneman et al. 2006; Batada and Hurst 2007). Depending on the interplay of mutational target size, robustness conferred by multiple regulators, and selective constraints on gene expression, in-degree might be either positively or negatively correlated with gene expression divergence.

The number of genes regulated by a transcription factor, a quantity known as 'out-degree', has also been predicted to influence the evolution of gene expression (McGuigan et al. 2014). Specifically, it has been proposed that mutations that alter expression of a transcription factor with many target genes should be more deleterious than mutations that alter expression of transcription factors with fewer target genes because the former have a greater potential to affect many phenotypes at once, increasing the probability that the mutation has deleterious effects on fitness (Cooper et al. 2007). Consistent with this

idea, studies examining the effects of individual gene deletions in the baker's yeast *Saccharomyces cerevisiae* have shown a significant negative correlation between the number of genes that change expression upon knockout of a gene and the relative fitness of that gene's deletion (Hughes et al. 2000; Featherstone and Broadie 2002). Simulations of regulatory evolution also show evidence of a negative correlation between out-degree and effects of gene deletions on fitness, but suggest that this correlation is quite weak (Siegal et al. 2007). Taken together, these data suggest that if a relationship is present between the out-degree of transcription factors and the evolution of gene expression, it should be negative, with transcription factors regulating more target genes showing less expression divergence among species than transcription factors regulating fewer target genes.

Here, we test these hypotheses about relationships between in-degree or out-degree and the evolution of gene expression by comparing a transcriptional regulatory network inferred for *Drosophila melanogaster* (Marbach et al. 2012) to expression differences observed within and between closely related *Drosophila* species (Coolon et al. 2014). Correlations between network topology and regulatory evolution are observed that suggest the architecture of existing transcriptional regulatory networks influences paths of future evolutionary change.

Results

Assessing reliability of the D. melanogaster transcriptional regulatory network

To examine the evolution of gene expression in the context of a transcriptional regulatory network, we used the "supervised" network that Marbach et al. (2012) constructed from

datasets describing conservation of transcription factor binding motifs, physical binding of transcription factors, chromatin marks, patterns of gene expression, and experimentally confirmed regulatory interactions curated in REDfly (Halfon et al. 2008). Statistically significant differences in expression within and between closely related *Drosophila* species were taken from Coolon et al. (2014), in which RNA-seq data were used to compare transcript abundance between African and North American strains of *D. melanogaster* (*mel-mel*), *Drosophila simulans* and *Drosophila sechellia* (*sim-sec*), and *D. melanogaster* and *D. simulans* (*mel-sim*). Differences in *cis*-regulatory activity between each pair of strains or species reported in Coolon et al. (2014) were also used to test for relationships between in-degree or out-degree and *cis*-regulatory evolution, as *cis*-regulatory activity might provide a more direct read-out of the relationship between transcription factors and their target genes. We restricted our analysis to the 4577 of 12,286 genes in the Marbach et al. (2012) regulatory network for which both expression differences and relative *cis*-regulatory activity were analyzed in all three comparisons (Coolon et al. 2014). Of these, 227 were transcription factors that appeared as regulators in the network and 4576 were target genes in the network; one transcription factor did not appear as a target gene in the network. The Coolon et al. (2014) and Marbach et al. (2012) datasets are described in more detail in the Materials and Methods section, and Figure 2.8 explains how these datasets were merged. A comparison of in-degree and out-degree for genes in the Marbach et al. (2012) network that were included and excluded in our study is shown in Figure 2.9.

Because the transcriptional regulatory network we used was derived from data collected from *D. melanogaster*, we first considered whether or not this network provided

a reasonable approximation of transcriptional regulatory networks in *D. sechellia* and *D. simulans*. These two species last shared a common ancestor with *D. melanogaster* ~2.5 million years ago (Cutter 2008), yet both can still form viable F1 hybrids with *D. melanogaster*, suggesting that their transcriptional regulatory networks remain largely compatible. The strong conservation of transcription factor binding sites between *D. melanogaster* and *D. yakuba* (Bradley et al. 2010), species which diverged twice as long ago as *D. melanogaster*, *D. simulans* and *D. sechellia* (Cutter 2008), further suggests that network topology should be largely conserved among the species examined.

If a transcriptional network reliably represents regulatory relationships, we expect that transcription factors in this network with altered expression should tend to have more target genes with altered expression than transcription factors with conserved expression. Indeed, for all three comparisons, we found that transcription factors with expression differences between the strains or species compared had a greater proportion of target genes with statistically significant expression differences than transcription factors without expression differences (Figure 2.1 A-C). We also expect the converse to be true: target genes with expression differences should be more likely to have regulators (transcription factors) with expression differences between the strains or species being compared than target genes with conserved expression. Again, the data analyzed were consistent with this expectation: the proportion of transcription factors with significant differences in expression between the strains or species compared was larger for target genes that showed significant differences in expression than for target genes that did not (Figure 2.1D-F). (An assessment of the sensitivity of this metric to errors in the network structure is presented in Supplementary Figure 2.3.)

In-degree correlates with differences in gene expression within and between species

As described in the Introduction, the number of transcription factors directly controlling a gene's expression (in-degree) has been predicted to correlate positively or negatively with gene expression divergence depending on the factor assumed to be primarily responsible for the correlation. To empirically determine the relationship between in-degree and expression divergence, we compared the in-degree distributions between genes with ($N_{mel\text{-}mel}$ = 1372, $N_{sim\text{-}sec}$ = 1281, $N_{mel\text{-}sim}$ = 1480) and without ($N_{mel\text{-}mel}$ = 3204, $N_{sim\text{-}sec}$ = 3295, $N_{mel\text{-}sim}$ = 3096) statistically significant expression differences between the strains and species examined (Figure 2.2). We found that the medians of the in-degree distributions for the two groups were significantly different for all three comparisons (Wilcoxon rank sum test, $P_{mel\text{-}mel} = 2 \times 10^{-14}$, $P_{sim\text{-}sec} = 2 \times 10^{-10}$, $P_{mel\text{-}sim} = 1 \times 10^{-12}$), with differentially expressed genes having a lower median in-degree than genes that were not differentially expressed (Figure 2.2A-C).

To better understand the relationship between in-degree and the evolution of gene expression, we asked how the proportion of genes with a significant expression difference changed with increasing in-degree. Consistent with the tendency for genes with an expression difference to have lower in-degree than genes without an expression difference, increasing in-degree was found to be associated with a decreasing proportion of genes with expression differences using logistic regression ($P_{mel\text{-}mel} < 2 \times 10^{-16}$, $P_{sim\text{-}sec} = 5 \times 10^{-9}$, $P_{mel\text{-}sim} = 9 \times 10^{-12}$, N = 4576 in all tests). We also compared in-degree of each gene to its magnitude of expression difference (regardless of whether or not this difference was statistically significant) and used the nonparametric Spearman's rank correlation

coefficient (ρ) to test for a significant relationship between the two (Figure 2.2D-F). This analysis showed that genes with larger in-degrees are not only less likely to have a significant expression difference between strains and species, but that the magnitude of any expression differences that do exist also tends to be smaller ($\rho_{mel\text{-}mel} = -0.17$ $P_{mel\text{-}mel} < 2x10^{-16}$; $\rho_{sim\text{-}sec} = -0.14$, $P_{sim\text{-}sec} < 2x10^{-16}$; $\rho_{mel\text{-}sim} = -0.17$, $P_{mel\text{-}sim} < 2x10^{-16}$; N = 4576 in all tests).

In-degree correlates with differences in cis-regulatory activity within and between species

To determine whether the relationship observed between in-degree and differences in transcript abundance (gene expression) also exists between in-degree and differences in *cis*-regulatory activity, we again divided genes into two groups, those with ($N_{mel\text{-}mel}$ = 316, $N_{sim\text{-}sec}$= 489, $N_{mel\text{-}sim}$ = 732) and without ($N_{mel\text{-}mel}$ = 4260, $N_{sim\text{-}sec}$ = 4087, $N_{mel\text{-}sim}$ = 3844) significant *cis*-regulatory differences, and compared their in-degree distributions. A significantly lower in-degree was observed for genes with differences in *cis*-regulatory activity using Wilcoxon rank sum tests to compare the medians of the in-degree distributions (Figure 2.3A-C, $P_{mel\text{-}mel} = 3x10^{-3}$, $P_{sim\text{-}sec} = 2x10^{-8}$, $P_{mel\text{-}sim} = 2x10^{-3}$). Logistic regressions also indicated that higher in-degree was associated with a decreased probability of differences in *cis*-regulatory activity between strains and species ($P_{mel\text{-}mel}$ = 0.003; $P_{sim\text{-}sec} = 2x10^{-10}$; $P_{mel\text{-}sim}$ = 0.0027; N = 4576 in all cases), and a significantly negative Spearman's rank correlation coefficient was observed between in-degree and the magnitude of differences in *cis*-regulatory activity (Figure 2.3D-F, $\rho_{mel\text{-}mel} = -0.08$, $P_{mel\text{-}mel} = 5x10^{-8}$; $\rho_{sim\text{-}sec} = -0.13$, $P_{sim\text{-}sec} < 2x10^{-16}$; $\rho_{mel\text{-}sim} = -0.07$, $P_{mel\text{-}sim} = 6x10^{-6}$, N = 4576 in

all cases). These findings suggest that the effects of in-degree on the evolution of *cis*-regulatory activity are at least partially responsible for the observed relationship between in-degree and differences in gene expression.

Out-degree correlates with differences in gene expression within but not between species

Expression of transcription factors with many target genes (higher out-degree) is often assumed to evolve more slowly than expression of transcription factors with fewer target genes (lower out-degree) because changing expression of the former is expected to have greater pleiotropic effects and hence greater selective constraint than changing the latter. To test this hypothesis, we compared the median out-degree between transcription factors with ($N_{mel\text{-}mel}$ = 65, $N_{sim\text{-}sec}$= 44, $N_{mel\text{-}sim}$ = 56) and without ($N_{mel\text{-}mel}$ = 162, $N_{sim\text{-}sec}$= 183, $N_{mel\text{-}sim}$ = 171) differences in expression in the *mel-mel*, *sim-sec*, and *mel-sim* comparisons using the same tests described above for in-degree. [Note that the smaller number of transcription factors (N=227) than target genes (N=4576) provides less power to detect similarly sized effects for out-degree than in-degree.] When comparing expression between two strains of *D. melanogaster*, we found evidence of the predicted patterns: lower out-degree for transcription factors with expression differences (Figure 2.4A, $P_{mel\text{-}mel} = 2x10^{-4}$) and fewer (logistic regression: $\beta = -0.001$, $P = 4x10^{-4}$, N = 227) as well as smaller (Figure 2.4D, Spearman's rank correlation: $\rho = -0.23$, $P = 5x10^{-4}$, N = 227) expression differences for transcription factors with higher out-degree. Surprisingly, these relationships were not seen in either of the interspecific comparisons. Rather, we found no statistically significant differences in median out-degree between transcription

factors with and without expression differences in the *sim-sec* and *mel-sim* comparisons (Figure 2.4B-C, $P_{sim\text{-}sec}$= 0.71, $P_{mel\text{-}sim}$ = 0.70; N = 227 in both cases) nor any significant correlation between the probability of expression differences and out-degree (logistic regression, $P_{sim\text{-}sec}$= 0.50, $P_{mel\text{-}sim}$ = 0.79, N = 227 in both cases) or the magnitude of expression differences and out-degree (Figure 2.4E-F, Spearman's rank correlation, $\rho_{sim\text{-}sec}$ = -0.094, $P_{sim\text{-}sec}$= 0.16; $\rho_{mel\text{-}sim}$ = -0.090, $P_{mel\text{-}sim}$ = 0.17, N = 227 in both cases). These results are especially surprising given that the effects of selection, which is assumed to be the force driving a negative correlation between out-degree and expression divergence, should be stronger between than within species.

The hypothesis that out-degree negatively correlates with expression divergence is based on the assumption that out-degree is a good proxy for pleiotropy; however, this assumption might not be true. To examine this possibility, we compared out-degree to the number of Gene Ontology categories associated with each transcription factor, a measure previously shown to be correlated with other empirical measures of pleiotropy in yeast (He and Zhang 2006). We found no significant correlation between out-degree and the number of Gene Ontology categories among the transcription factors examined (Figure 2.11, Spearman's rank correlation coefficient = -0.07, P = 0.4). We also tested whether the number of Gene Ontology terms associated with a transcription factor correlates significantly with expression differences within or between species and found evidence for such a correlation only in the *mel-mel* comparison (Figure 2.5A-C, $P_{mel\text{-}mel}$ = 0.02 ; $P_{sim\text{-}sec}$ = 0.11; $P_{mel\text{-}sim}$ = 0.17). Similarly, the number of Gene Ontology terms only showed a statistically significant correlation with the magnitude of expression differences

in the *mel-mel* comparison (Figure 2.5D-F, $\rho_{mel\text{-}mel} = 0.16$, $P_{mel\text{-}mel} = 0.05$; $\rho_{sim\text{-}sec} = -0.05$, $P_{sim\text{-}sec} = 0.52$; $\rho_{mel\text{-}sim} = 0.00$, $P_{mel\text{-}sim} = 0.95$, $N = 227$ in all cases). In both of these cases, however, the significant relationship observed in the *mel-mel* comparison between the number of Gene Ontology terms and expression differences was in the opposite direction than expected, with transcription factors having more ontology terms more likely to have an expression difference (Figure 2.5A) or a larger expression difference (Figure 2.5D) than transcription factors with fewer ontology terms.

Out-degree does not correlate with differences in cis-regulation within or between species

To determine whether the relationship between out-degree and expression differences seen for the *mel-mel* comparison might be explained by a correlation between out-degree and differences in *cis*-regulatory activity, we compared out-degree between transcription factors with ($N_{mel\text{-}mel} = 10$) and without ($N_{mel\text{-}mel} = 217$) *cis*-regulatory differences in the *mel-mel* comparison. We found no significant difference in the median out-degree between the two groups of genes (Figure 2.6A, $P = 0.71$) nor any significant correlation between out-degree and the probability of *cis*-regulatory differences (logistic regression, $P = 0.57$, $N_{mel\text{-}mel} = 227$) or magnitude of *cis*-regulatory differences (Figure 2.6D, Spearman's $\rho = -0.01$, $P = 0.88$, $N_{mel\text{-}mel} = 227$). For completeness, we also tested *cis*-regulatory differences in the *sim-sec* and *mel-sim* comparisons for a correlation with out-degree. Again, we found no significant difference in out-degree between transcription factors with ($N_{sim\text{-}sec} = 16$, $N_{mel\text{-}sim} = 22$) and without ($N_{sim\text{-}sec} = 211$, $N_{mel\text{-}sim} = 205$) differences in *cis*-regulatory activity (Figure 2.6B-C, $P_{sim\text{-}sec} = 0.56$, $P_{mel\text{-}sim} = 0.44$) nor

any significant correlation between out-degree and the probability of *cis*-regulatory differences (logistic regression, $P_{sim\text{-}sec} = 0.27$; $P_{mel\text{-}sim} = 0.38$; N = 227 in both cases) or magnitude of *cis*-regulatory differences (Figure 2.6E-F, Spearman's $\rho_{sim\text{-}sec} = -0.08$, $P_{sim\text{-}sec} = 0.21$; $\rho_{mel\text{-}sim} = -0.07$, $P_{mel\text{-}sim} = 0.32$, N = 227 in both cases). These data suggest that the out-degree of a transcription factor has little influence on the evolution of its *cis*-regulatory expression differences.

Discussion

By comparing in-degree and out-degree in the *Drosophila* regulatory network with changes in gene expression and *cis*-regulation that have evolved within and between species, we found that genes regulated by larger numbers of transcription factors tended to have fewer and smaller changes in expression both within and between species than genes regulated by smaller numbers of transcription factors. By contrast, we found that the number of genes a transcription factor regulates, a property predicted to be related to pleiotropy, showed a statistically significant correlation with differences in total gene expression only when comparing two strains of *D. melanogaster*. No significant correlation between out-degree and differences in *cis*-regulation were observed in any comparison, either within or between species. These relationships are summarized in Figure 2.7. Below, we discuss the implications of our findings and compare our results with results from a similar study of regulatory differences between *S. cerevisiae* and *S. paradox* (Kopp and McIntyre 2012).

Network in-degree appears to influence the evolution of gene expression

The combinatorial control of a gene's expression by sets of transcription factors might either suppress or enhance the effects of new mutations on transcript levels. For example, genes regulated by many transcription factors may be more likely to have their expression altered by new mutations than genes regulated by fewer transcription factors because there are more sites in the genome that affect the expression of these genes. Consistent with this prediction, a study of mutation accumulation lines in yeast found that mutational variance (differences in gene expression caused by new mutations) correlated positively with the number of *trans*-acting regulators predicted to regulate a gene's expression (Landry et al. 2007). Interactions among transcription factors regulating expression of a gene can complicate the relationship between mutational target size and changes in gene expression, however. For example, a mutation that disrupts activity of a transcription factor might have little to no effect on expression of a target gene if another transcription factor(s) partially or completely compensates for the loss of the first transcription factor's activity. Effects of mutating individual transcription factors might also be smaller for genes regulated by larger sets of transcription factors than smaller sets if each transcription factor contributes a comparable amount to gene expression. These properties might cause genes regulated by large sets of transcription factors to acquire changes in expression more slowly and/or less often than genes regulated by fewer transcription factors. Our data are consistent with these latter models: the number of transcription factors regulating a gene's expression showed a significant negative correlation with both the frequency and magnitude of total expression differences as well as *cis*-regulatory differences within and between species.

A similar comparison between the number of transcription factors regulating a gene's expression and its *cis*-regulatory divergence was performed for *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* by Kopp and McIntyre (2012). They found the opposite relationship between network in-degree and *cis*-regulatory divergence, with larger differences in *cis*-regulation observed for genes with larger numbers of transcriptional regulators, but the magnitude of this effect was described as small. The different relationships observed in these two studies might result from differences in the structure of transcriptional regulatory networks between yeast and flies. For example, compared to genes in *Drosophila*, *Saccharomyces* genes tend to have relatively few regulators (Figure 2.12), which limits the opportunity for interactions among transcription factors to buffer the effects of regulatory changes. The smaller number of regulators might also cause genetic changes affecting a single transcriptional regulator to tend to have larger effects on expression. Ultimately, however, the reason for the different relationships reported between in-degree and regulatory divergence in yeast (Kopp and McIntyre 2012) and flies (this study) remains unknown.

Network out-degree appears to have minimal effect on the evolution of gene expression

Patterns of variation within and between species are influenced by both mutation and selection, with selection acting to preserve favorable genetic variants and eliminate deleterious ones. In the context of regulatory networks, genetic variants that affect expression of genes that influence activity of many other genes are thought to often be deleterious because of their expected greater pleiotropy and thus expected to be preferentially eliminated by natural selection (e.g., Featherstone and Broadie 2002;

Cooper et al. 2007; McGuigan et al. 2014). Our data provide limited support for this hypothesis, however, with the predicted negative correlation between divergence and the number of target genes of a transcription factor observed only for expression differences within a species. Selection is expected to have a larger impact on differences between than within species because of the longer divergence time, suggesting that selection is unlikely to be responsible for the relationship observed within *D. melanogaster*. Kopp and Mclntyre (2012) also failed to find a statistically significant correlation between out-degree of a transcription factor and *cis*-regulatory divergence between *S. cerevisiae* and *S. paradoxus.*

We do not think that these findings refute the idea that increasing pleiotropy increases the probability that a genetic change is deleterious, but rather that they suggest that the number of direct target genes a transcription factor regulates ("out-degree") is not a good measure of pleiotropy. For example, the target genes of a transcription factor might tend to affect the same biological functions, minimizing the pleiotropic effects of genetic changes affecting expression of that transcription factor. The absence of a correlation between out-degree and the number of Gene Ontology terms associated with a transcription factor is consistent with this potential explanation. The number of Gene Ontology terms also failed to correlate with expression divergence, however, suggesting that it might also be a poor measure of pleiotropy (at least in Drosophila). Quantifying pleiotropy is notoriously difficult (Paaby and Rockman 2013), and detecting any relationship between pleiotropy and the evolution of gene expression that might exist will likely require information beyond the topology of regulatory networks and Gene Ontology categorizations.

Looking ahead

Understanding how existing biological systems shape the paths for future evolutionary change is an important goal for evolutionary biology. We must understand how genotypes are translated into phenotypes to achieve this goal, and the elucidation of regulatory networks controlling gene expression is a key step in this process. Our results suggest that some topological features of regulatory networks (e.g., in-degree) might be useful predictors of evolutionary change, whereas others (e.g., out-degree) might have less explanatory power than expected. The scope of these conclusions is limited, however, by the small number of species for which even a preliminary comparison between network topology and expression divergence is possible; elucidating regulatory networks remains challenging in even the most developed genetic model systems. Advances in functional genomics, computational tools for inferring regulatory networks, and methods for perturbing genomes to assess the phenotypic effects of a particular genetic change promise to provide more opportunities to study the relationship between biological networks and regulatory evolution.

Materials and Methods

Transcriptional regulatory network

The transcriptional regulatory network used in this work was the "supervised" network described in Marbach et al. (2012). It was inferred using information from several sources, including genome-wide chromatin immuno-precipitation, conserved transcription factor binding motifs among 12 *Drosophila* species, gene expression

profiles across different development stages, chromatin modification profiles among several cell types, and experimentally confirmed regulatory relationships (Marbach et al. 2012).

Additional tests of the reliability of this network and its applicability to other Drosophila species (*D. simulans* and *D. sechellia*) were performed by switching edges among genes in the network as shown in Supplementary Figure 2.3. Although this approach is intuitive and has been used to compare observed and randomized network structures in prior work (e.g., Milo et al. 2002, 2003; Iorio et al. 2016), the statistical properties of the null models generated in this way have not been established for comparing to datasets like gene expression with covariance among measures and the results should be interpreted with this in mind (Churchill and Doerge 2008). Briefly, the degree-preserving network randomization was done by randomly selecting two edges in the network and then, as long as the newly created edges did not already exist in the network, exchanging their target genes. This process was repeated until the intended percentage of edges was switched. In other words, 10% edge switching means 10% of the edges have exchanged ends with other edges. This randomization strategy keeps the in-degree and out-degree unchanged for all randomized networks (Milo et al. 2002, 2003). Error bars shown in Supplementary Figure 2.3B-G indicate two standard deviations around the mean derived from the 200 permutations with the same percent edge switching.

Comparing gene expression and cis-regulatory activity among strains and species

Differences in mRNA transcript abundance ("gene expression") and relative *cis*-regulatory activity between the *zhr* and *z30* strains of *D. melanogaster*, the *droSec1* strain of *D. sechellia* and Tsimbazaza strain of *D. simulans*, and the *zhr* strain of *D. melanogaster* and Tsimbazaza strain of *D. simulans* were taken from the analysis of RNA-seq data described in Coolon et al. (2014). These data include comparisons of gene expression between each pair of strains or species (*mel-mel*, *sim-sec*, and *mel-sim*) as well as comparisons of relative *cis*-regulatory activity inferred by comparing relative allelic expression in F1 hybrids produced by crossing each pair of strains or species (Wittkopp et al. 2004; McManus et al. 2010). The statistical significance of differences in gene expression and *cis*-regulatory activity between strains or species were determined using binomial exact tests to compare read abundance in mixed parental (for expression differences) and F1 hybrid (for *cis*-regulatory differences) RNA-seq datasets with a Benjamini and Hochberg (1995) 5% false discovery rate (as implemented in R v3.0.1) to correct for multiple testing (Coolon et al. 2014). The process used to merge the expression and network files and identify the 4577 genes analyzed in this study is described in Figure 2.8. Ultimately, we analyzed the 4577 genes (including 227 transcription factors) that passed the quality control standards for measuring allele-specific expression used by Coolon et al. (2014) and also appeared in the regulatory network inferred by Marbach et al. (2012) (Figure 2.8). All gene annotations were based on *D. melanogaster* FlyBase FBgn#s (Attrill et al. 2015).

Comparing network properties to differences in gene expression and cis-regulation

Analyses shown in Figures 2.1, 2.2A-C, 2.3A-C, 2.4A-C, and 2.6A-C compare the presence or absence of statistically significant (FDR = 0.05) differences in gene expression or *cis*-regulatory activity described in Coolon et al. (2014) to relationships among genes in the network (Figure 2.1), in-degree of all target genes (Figures 2.2 and 2.3) and out-degree of all transcription factors (Figure 2.4 and 2.6). Non-parametric Wilcoxon rank sum tests were used to compare median in-degree and out-degree between sets of genes with and without statistically significant differences in gene expression or *cis*-regulation for each pair of strains or species examined as well as to compare the proportion of target genes with differential expression between transcription factors with and without differential expression and vice versa. These tests evaluated the null hypothesis of no association between in-degree or out-degree and differences in gene expression or *cis*-regulation. Logistic regressions were also used to compare an indicator variable representing whether or not a gene had a statistically significant difference in gene expression and/or *cis*-regulatory activity in a given comparison to its in-degree or out-degree. These tests were performed using the glm function in R with the options "family=binomial, link=logit", which uses a Z-score to assess the statistical significance of the factor being tested; a significant test indicates that the factor tested (e.g., in-degree or out-degree) has statistically significant predictive ability for which genes have significant expression differences. The null hypothesis in each case was that the factor tested was not a significant predictor of differences in expression or *cis*-regulation.

Spearman's rank correlation coefficients were used to test for a significant correlation between the $\log_2$ transformed magnitude of the differences in gene expression or *cis*-

regulatory activity reported in Coolon et al. (2014) and a gene's in-degree or out-degree. The null hypothesis for this test is that there is no relationship between a gene's in-degree or out-degree and the magnitude of its expression difference between strains or species. Results from these tests are shown in Figures 2.2D-F, 2.3D-F, 2.4D-F, and 2.6D-F. A LOESS (locally weighted scatterplot smoothing**)** line was fitted to these data using the loess function with default parameters in R.

Gene Ontology (GO) analysis

Gene Ontology terms were obtained from FlyBase (Attrill et al. 2015) for each transcription factor in our dataset, and the number of GO terms associated with each transcription factor was used as a proxy for its degree of pleiotropy. To minimize redundancy among GO terms, we restricted our analysis to the GO SLIM categories defined by Gene Ontology Consortium (2015). To determine whether the number of GO SLIM terms associated with a transcription factor was related to differences in expression of its target genes between each pair of strains or species, we used Wilcoxon rank sum tests to compare the median number of GO SLIM terms between sets of transcription factors with and without expression differences (Figure 2.5A-C). Spearman's rank correlation coefficients were also used to test for a significant relationship between the $\log_2$ magnitude of differences in gene expression or *cis*-regulatory activity and number of GO SLIM terms associated with each transcription factor.

Comparing in-degree distributions between flies and yeast

In the Discussion section, we compare our results from analysis of a Drosophila regulatory network to a similar study that was performed using a *S. cerevisiae* regulatory network (Kopp and McIntyre 2012), including a comparison of the in-degree distributions between the two networks. The *Drosophila melanogaster* network used for this analysis was the same network used for the rest of the analyses in this paper (Marbach et al. 2012) and the *S. cerevisiae* network used was described in Balaji et al. (2006). In each case, in-degree was calculated as the number of transcription factors predicted to regulate a target within the network.

Statistical analyses

All statistical analyses were performed in R v3.2.2 (RCoreTeam 2016). Database files and scripts used to perform these analyses are available for download from https://deepblue.lib.umich.edu/data/concern/generic_works/9s161628x. Supplementary Figures 2.1 and 2.3 and their associated legends describe which files were used for each step of the project.

Acknowledgements

We thank the University of Michigan LSA High Performance Computing for computational resources; Joseph Coolon for providing the gene expression datasets; Joseph Coolon, Kraig Stevenson and Brian Metzger for advice on statistical analysis; and Brian Metzger and Fabien Duveau for comments on the manuscript. Funding for this work was from the National Science Foundation (MCB-1021398).

References

Attrill H, Falls K, Goodman JL, Millburn GH, Antonazzo G, Rey AJ, Marygold SJ, FlyBase Consortium. 2015. FlyBase: establishing a Gene Group resource for Drosophila melanogaster. Nucleic Acids Res. 44:D786–D792.

Balaji S, Babu MM, Iyer LM, Luscombe NM, Aravind L. 2006. Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. J. Mol. Biol. 360:213–227.

Batada NN, Hurst LD. 2007. Evolution of chromosome organization driven by selection for reduced gene expression noise. Nat Genet 39:945–949.

Benjamini, Y, Hochberg, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Series B Stat Methodol 57: 289-300.

Borneman AR, Leigh-Bell JA, Yu H, Bertone P, Gerstein M, Snyder M. 2006. Target hub proteins serve as master regulators of development in yeast. Genes Dev. 20:435–448.

Bradley RK, Li X-Y, Trapnell C, Davidson S, Pachter L, Chu HC, Tonkin LA, Biggin MD, Eisen MB. 2010. Binding Site Turnover Produces Pervasive Quantitative Changes in Transcription Factor Binding between Closely Related Drosophila Species. PLOS Biol 8:e1000343.

Carroll SB. 2008. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. Cell 134:25–36.

Churchill GA, Doerge RW. 2008. Naive application of permutation testing leads to inflated type I error rates. Genetics 178: 609–610.

Coolon JD, McManus CJ, Stevenson KR, Graveley BR, Wittkopp PJ. 2014. Tempo and mode of regulatory evolution in Drosophila. Genome Res. 24:797–808.

Cooper TF, Ostrowski EA, Travisano M. 2007. A negative relationship between mutation pleiotropy and fitness effect in yeast. Evolution 61: 1495–1499.

Cutter AD. 2008. Divergence times in Caenorhabditis and Drosophila inferred from direct estimates of the neutral mutation rate. Mol. Biol. Evol. 25:778–786.

Evangelisti AM, Wagner A. 2004. Molecular evolution in the yeast transcriptional

regulation network. J Exp Zool Part B 302B:392–411.

Featherstone DE, Broadie K. 2002. Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network. Bioessays 24:267–274.

Gene Ontology Consortium. 2015. Gene Ontology consortium: going forward. Nucleic Acids Res. 43:D1049-D1056.

Halfon MS, Gallo SM, Bergman CM. 2008. REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in Drosophila. Nucl. Acids Res. 36:D594–D598.

He X, Zhang J. 2006. Toward a molecular understanding of pleiotropy. Genetics 173:1885–1891.

Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, et al. 2000. Functional Discovery via a Compendium of Expression Profiles. Cell 102:109–126.

Iorio F, Bernardo-Faura M, Gobbi A, Cokelaer T, Jurman G, Saez-Rodriguez J. 2016. Efficient randomization of biological networks while preserving functional characterization of individual nodes. BMC Bioinformatics 17: 542.555.

Jovelin R, Phillips PC. 2009. Evolutionary rates and centrality in the yeast gene regulatory network. Genome Biol. 10:R35.

Kopp A, McIntyre LM. 2012. Transcriptional network structure has little effect on the rate of regulatory evolution in yeast. Mol. Biol. Evol. 29:1899–1905.

Landry CR, Lemos B, Rifkin SA, Dickinson WJ, Hartl DL. 2007. Genetic Properties Influencing the Evolvability of Gene Expression. Science 317:118–121.

Macneil LT, Walhout AJM. 2011. Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. Genome Res. 21:645–657.

Marbach D, Roy S, Ay F, Meyer PE, Candeias R, Kahveci T, Bristow CA, Kellis M. 2012. Predictive regulatory models in Drosophila melanogaster by integrative inference of transcriptional networks. Genome Res. 22:1334–1349.

McGuigan K, Collet JM, Allen SL, Chenoweth SF, Blows MW. 2014. Pleiotropic Mutations Are Subject to Strong Stabilizing Selection. Genetics 197:1051–1062.

McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp PJ. 2010. Regulatory divergence in Drosophila revealed by mRNA-seq. Genome Res. 20:816–

825.

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. 2002. Network motifs: simple building blocks of complex networks. Science 298:824–827.

Milo R, Kashtan N, Itzkovitz S, Newman, MEJ, Alon U. 2003. On the uniform generation of random graphs with prescribed degree sequences. arXiv:cond-mat/0312028

Paaby AB, Rockman MV. 2013. The many faces of pleiotropy. Trend Genet 29:66–73.

Promislow D. 2005. A Regulatory Network Analysis of Phenotypic Plasticity in Yeast. Am Nat 165:515–523.

RCoreTeam. 2016. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Siegal ML, Promislow DEL, Bergman A. 2007. Functional and evolutionary inference in gene networks: does topology matter? Genetica 129:83–103.

Stern DL, Orgogozo V. 2009. Is Genetic Evolution Predictable? Science 323:746–751.

Wittkopp PJ, Haerum BK, Clark AG. 2004. Evolutionary changes in cis and trans gene regulation. Nature 430:85-88.

Wittkopp PJ, Kalay G. 2011. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat Rev Genet 13:59-69.

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 8:206–216.

Zhu X, Gerstein M, Snyder M. 2007. Getting connected: analysis and principles of biological networks. Genes Dev. 21:1010–1024.

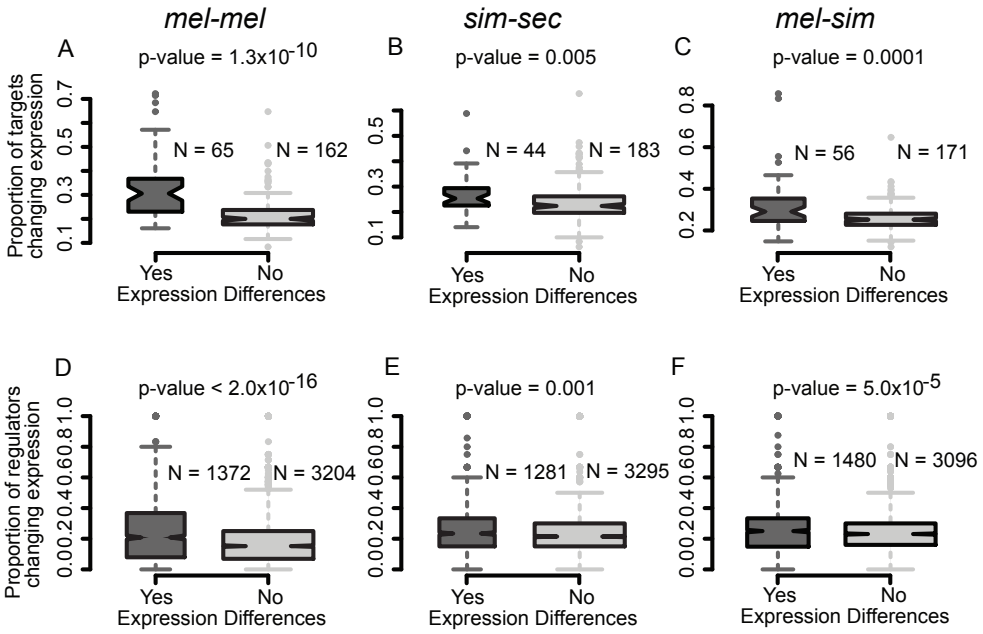

Figure 2.1. Assessing reliability of the regulatory network. (**A-C**) For each transcription factor (N = 227), we calculated the proportion of its target genes that showed significant expression differences between the strains or species compared. The boxplots show the distributions of these proportions for transcription factors with (dark grey) and without (light grey) significant expression differences between the two strains of *D. melanogaster* (**A**), *D. simulans* and *D. sechellia* (**B**), and *D. melanogaster* and *D. simulans* (**C**). P-values shown are from non-parametric Wilcoxon rank sum tests, and N indicates the number of transcription factors included in each category. (**D-F**) For each target gene (N = 4576), we calculated the proportion of regulators (transcription factors) that showed significant expression differences between the strains or species being compared. The boxplots show the distributions of these proportions for target genes with (dark grey) and without (light grey) significant expression differences between the two strains of *D. melanogaster* (**D**), *D. simulans* and *D. sechellia* (**E**), and *D. melanogaster* and *D. simulans* (**F**). P-values shown are from non-parametric Wilcoxon rank sum tests, and N indicates the number of target genes included in each category.

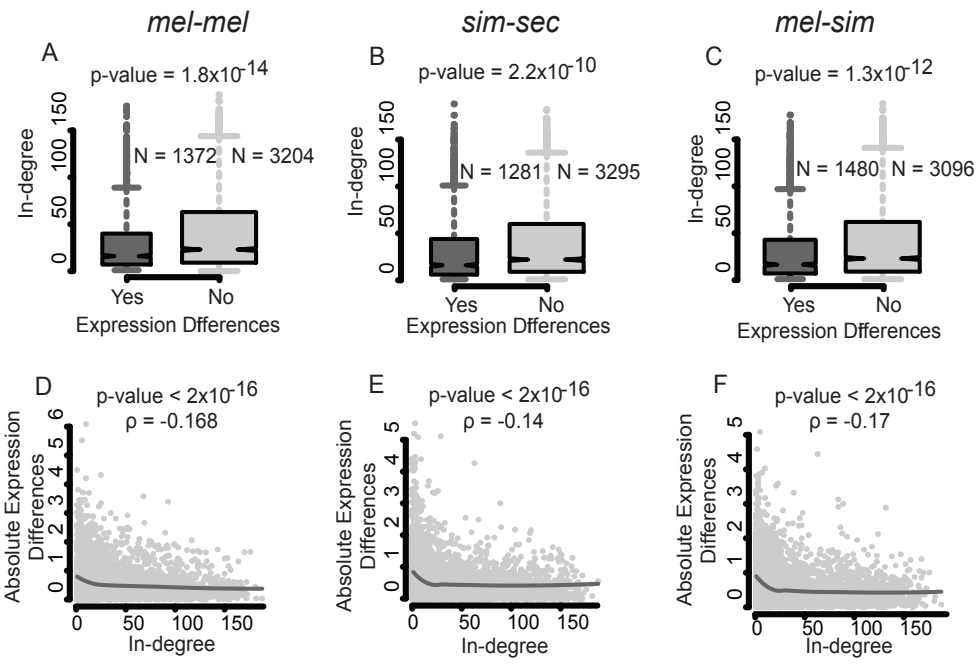

Figure 2.2. Relationship between network in-degree and differences in gene expression within species and between species. (**A-C**) Boxplots show the in-degree distributions for genes with (dark grey) and without (light grey) significant differences in gene expression in the *mel-mel* (**A**), *sim-sec* (**B**), and *mel-sim* (**C**) comparisons. P-values are from non-parametric Wilcoxon rank sum tests, and N indicates the number of genes in each group. (**D-F**) Absolute magnitude of gene expression differences (Y-axis) is plotted against in-degree (X-axis) in the *mel-mel* (**D**), *sim-sec* (**E**), and *mel-sim* (**F**) comparisons. A LOESS line fitted to these data is shown in dark grey. Spearman's rank correlation coefficients (ρ) and associated p-values are also shown.

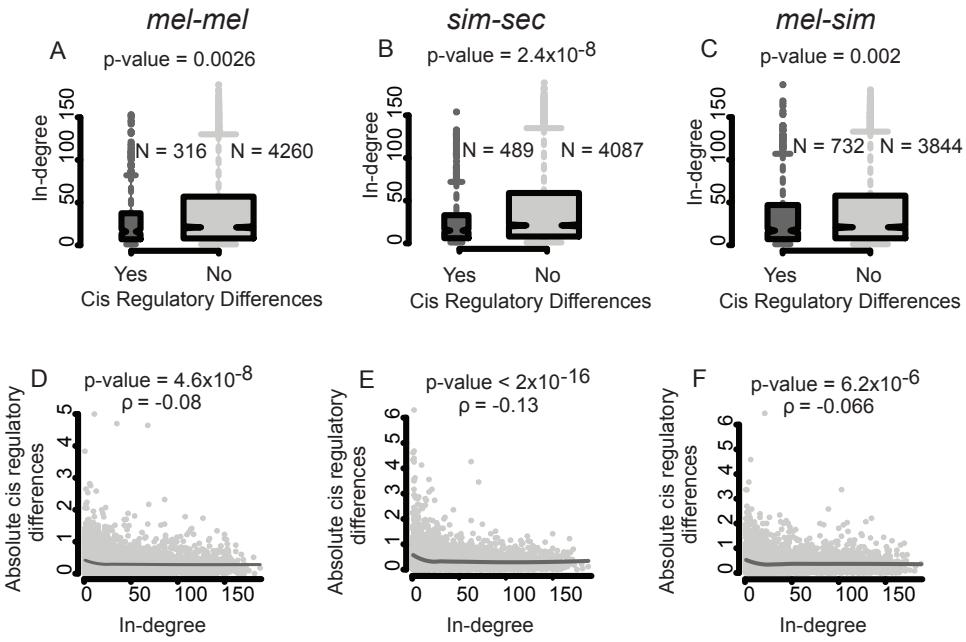

Figure 2.3. Relationship between network in-degree and difference in *cis*-regulatory activity within species and between species. (**A-C**) Boxplots show in-degree distributions for genes with (dark grey) and without (light grey) significant differences in *cis*-regulatory activity in the *mel-mel* (**A**), *sim-sec* (**B**), and *mel-sim* (**C**) comparisons. P-values are from non-parametric Wilcoxon rank sum tests, and N indicates the number of genes in each group. **(D-F**) Absolute magnitude of *cis*-regulatory difference (Y-axis) is plotted against in-degree (X-axis) in the *mel-mel* (**D**), *sim-sec* (**E**), and *mel-sim* (**F**) comparisons. A LOESS line fitted to these data is shown in dark grey. Spearman's rank correlation coefficients (ρ) and associated p-values are also shown.

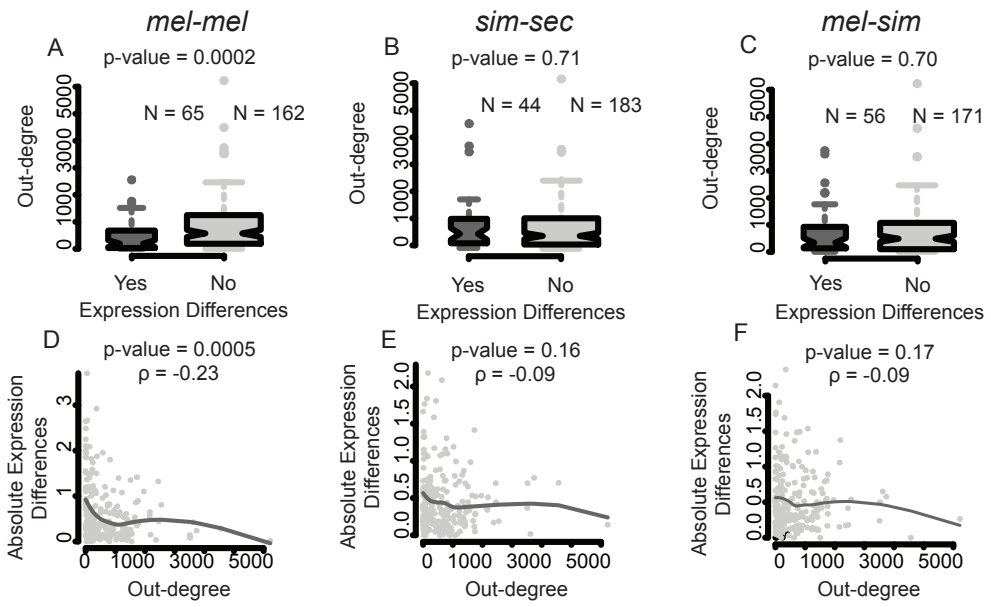

Figure 2.4. Relationship between network out-degree and difference in gene expression within species and between species. (**A-C**) Boxplots show out-degree distributions for genes with (dark grey) and without (light grey) significant differences in gene expression in the *mel-mel* (**A**), *sim-sec* (**B**), and *mel-sim* (**C**) comparisons. P-values are from non-parametric Wilcoxon rank sum tests, and N indicates the number of genes in each group. (**D-F**) Absolute magnitude of gene expression difference (Y-axis) is plotted against out-degree (X-axis) in the *mel-mel* (**D**), *sim-sec* (**E**), and *mel-sim* (**F**) comparisons**.** A LOESS line fitted to these data is shown in dark grey. Spearman's rank correlation coefficients (ρ) and associated p-values are also shown.

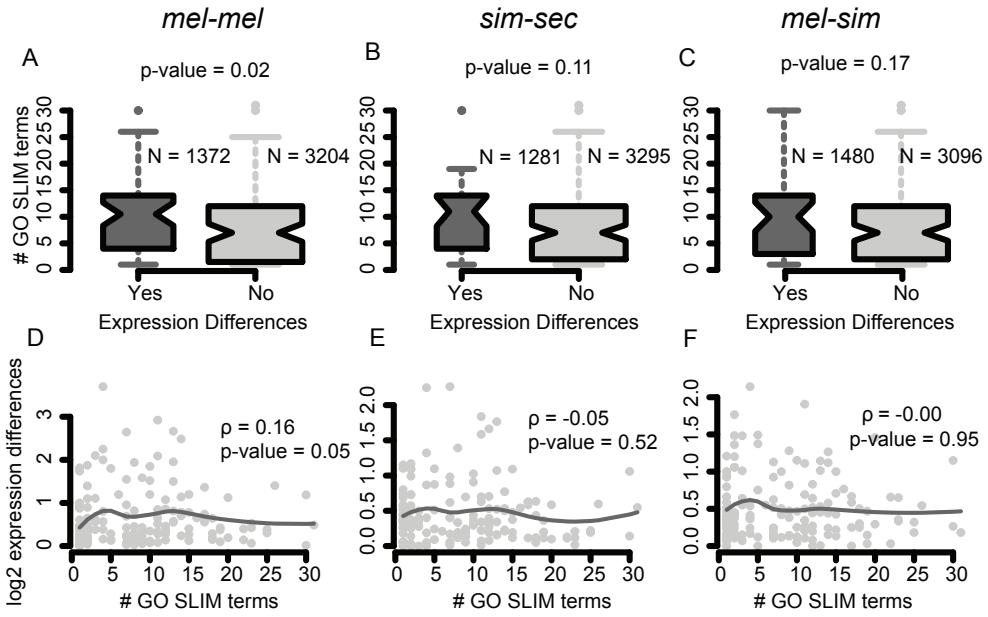

Figure 2.5. Relationship between number of GO SLIM terms associated with a transcription factor and differences in gene expression within species and between species. (**A-C**) Boxplots show GO SLIM term distributions for genes with (dark grey) and without (light grey) significant differences in gene expression in the *mel-mel* (**A**), *sim-sec* (**B**), and *mel-sim* (**C**) comparisons. P-values are from non-parametric Wilcoxon rank sum tests, and N indicates the number of genes in each group. (**D-F**) Absolute magnitude of gene expression differences (Y-axis) is plotted against the number of GO SLIM terms (X-axis) in the *mel-mel* (**D**), *sim-sec* (**E**), and *mel-sim* (**F**) comparisons. A LOESS line fitted to these data is shown in dark grey. Spearman's rank correlation coefficients (ρ) and associated p-values are also shown.

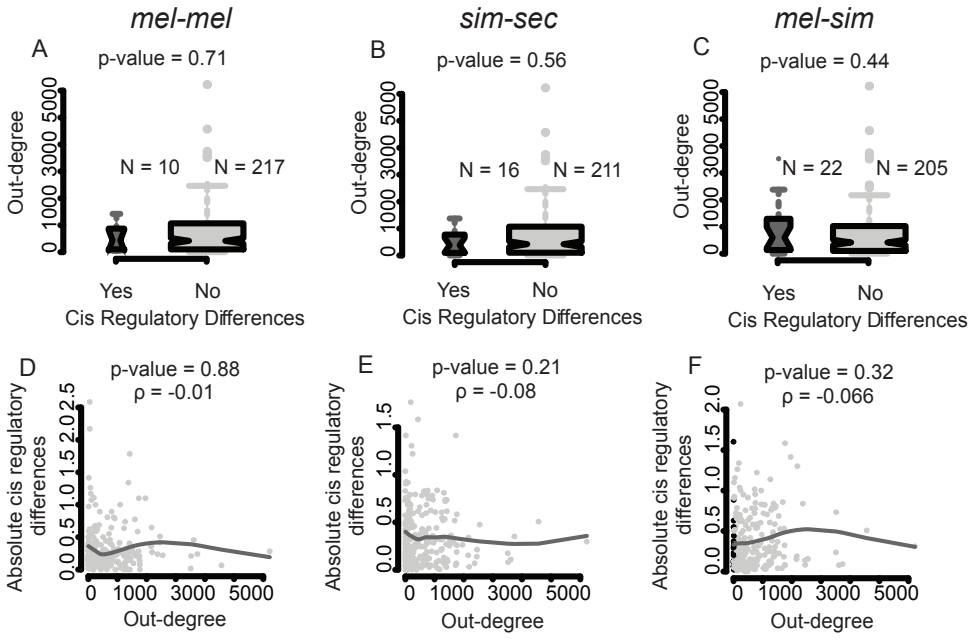

Figure 2.6. Relationship between network out-degree and difference in *cis*-regulatory activity within species and between species. (**A-C**) Boxplots show out-degree distributions for genes with (dark grey) and without (light grey) significant differences in *cis*-regulation in the *mel-mel* (**A**), *sim-sec* (**B**), and *mel-sim* (**C**) comparisons. P-values are from non-parametric Wilcoxon rank sum tests, and N indicates the number of genes in each group. (**D-F**) Absolute magnitude of *cis*-regulatory differences (Y-axis) is plotted against out-degree (X-axis) in the *mel-mel* (**D**), *sim-sec* (**E**), and *mel-sim* (**F**) comparisons. A LOESS line fitted to these data is shown in dark grey. Spearman's rank correlation coefficients (ρ) and associated p-values are also shown.

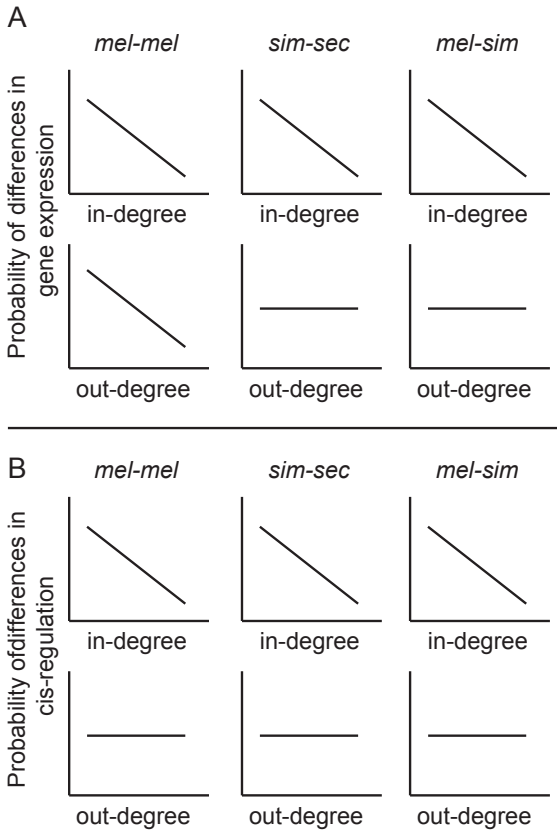

Figure 2.7. In-degree is a better predictor of changes in *cis*-regulation and gene expression over evolutionary time than out-degree. This schematic shows the direction of the relationship, if any, between differences in gene expression (A) or *cis*-regulation (B) observed within or between Drosophila species and either in-degree (top row) or out-degree (bottom row), which are properties of the network architecture. A horizontal line indicates that no statistically significant relationship (defined as $P < 0.01$ for the Wilcoxon rank sum test) was observed. As described in the main text and shown in Figure 2.5, we also compared differences in transcription factor expression to the number of GOSlim terms associated with each transcription factor and found evidence of a marginally significant relationship ($p = 0.02$ for Wilcoxon test, $p = 0.05$ for Spearman's rank correlation coefficient only in the *mel-mel* comparison and the sign of this correlation was in the opposite direction than predicted.

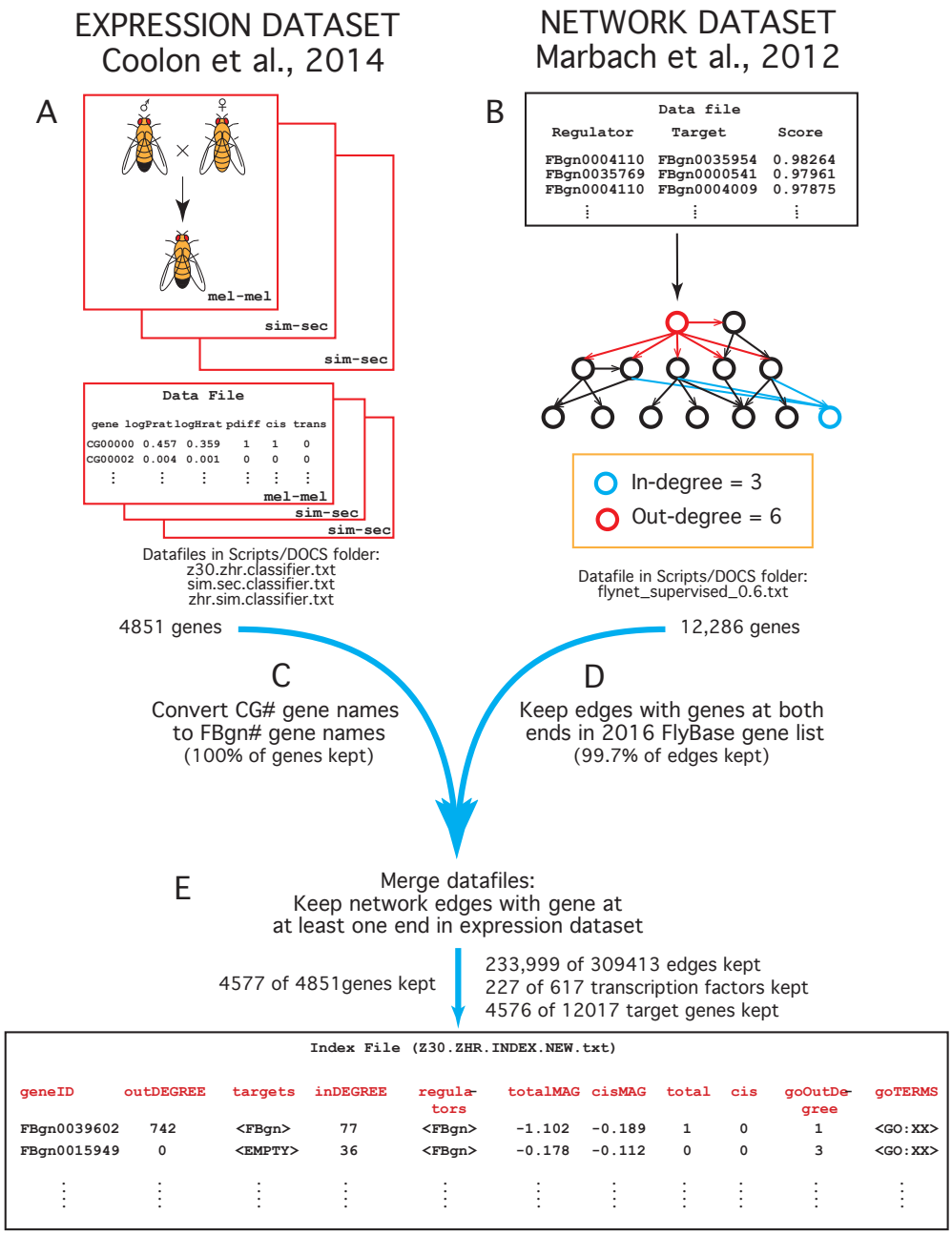

Figure 2.8. Integrating network structure and expression divergence. (**A**) Differences in gene expression and *cis*-regulation between strains and species of Drosophila were derived from RNAseq data collected from adult females of each genotype and F1 hybrids produced by crossing each pair of strains or species, as

described in Methods. The difference in expression between each pair of ("parental") strains or species was reported for each **gene** as $\log_2$(genotype 1 read count/genotype 2 read count) with the magnitude of this ratio shown in the **LogPrat** column in the data file. The statistical significance of any difference in expression level between the pair of genotypes was determined using a binomial exact test followed by a Benjamini-Hochberg false discovery rate correction for multiple testing as described in Methods. This significance is indicated in the **pdiff** column in the datafile, with 1 = significant, 0 = nonsignificant. The *cis*-regulatory difference for each gene was reported as $\log_2$(allele 1 read count/allele 2 read count) in the F_1 hybrid. The magnitude of this difference is shown in the **LogHrat** column in the data file. The statistical significance of any difference in *cis*-regulatory activity between the two alleles was determined using a binomial exact test followed by a Benjamini-Hochberg false discovery rate correction for multiple testing as described in Methods. This significance is indicated in the **cis** column in the datafile, with 1 = significant and 0 = nonsignificant. The **trans** column in this datafile was not used in our analyses. These datafiles were downloaded from the supplementary materials of Coolon et al. (2014) and are named z30.zhr.classifier.txt, sim.sec.classifier.txt, and zhr.sim.classifier.txt. (**B**) The data file describing the network used in this work (flynet_supervised_0.6.txt) was downloaded from the supplementary materials of Marbach et al. (2012). For each transcription factor (**Regulator**) - target gene (**Target**) pair, the confidence score (**Score**) describes the probability of the edge calculated by the Marbach et al. (2012) supervised method. All edges with a probability >0.6 were retained in the network used for our work, which is the same cutoff used by Marbach et al. (2012) for their analyses. A sample network is shown along with the in-degree (# of regulators a target has) of the blue node and out-degree (# of targets a regulator has) of the red node. (**C**) To prepare to merge the gene expression and network datafiles, we converted the CG Gene ID numbers in the expression data to the FBgn ID numbers (**Primary FBgn#**) in the fbgn_annotation_ID-fb_2016_01.tsv file downloaded from FlyBase. No genes were eliminated at this step; all genes in the Coolon et al. (2014) datafile had a corresponding FBgn#. (**D**) FBgn#s from the network datafile were compared to the 2016 FBgn gene list from FlyBase (fbgn_annotation_ID-fb_2016_01.tsv). Edges in the network datafile were excluded if either the regulator or target did not have a corresponding **Primary FBgn#** in this datafile, eliminating 0.3% of edges from the original Marbach et al. (2012) network. (**E**) The expression and network datafiles were then merged using the INFOPROCESSING.py script, run using the COMMAND.sh script. 4577 of 4851 (94.4%) genes in the expression datafile were kept because they appeared at least once in the network. Edges in the network file were kept if the regulator and/or target gene was present in the expression dataset. This filtering retained 233,999 of 309,413 (75.6%) edges, 227 of 617 (36.8%)

regulators, and 4,576 of 12,017 (28.1%) targets. As shown in Figure 2.9, most regulators excluded had very few targets and most targets excluded had very few regulators, explaining why we excluded less than 25% of edges while excluding over 60% of all genes. The merged datafiles are provided as OUTPUT/z30.zhr.INDEX.NEW.txt, OUTPUT/sim.sec.INDEX.NEW.txt, and OUTPUT/zhr.sim.INDEX.NEW.txt, with the format of these files shown. Column names in these files are defined as follows: **geneID**: Flybase gene name. **outDEGREE**: number of target genes for each regulator (transcription factor); set to zero if the gene is not a transcription factor. **targets**: name(s) of target genes; **InDEGREE**: number of transcription factors putatively regulating each gene. **regulators**: name(s) of regulators. **totalMAG**: magnitude of total expression difference between strains or species; derived from logPrat in expression file. **cisMag**: magnitude of *cis*-regulatory difference; derived from logHrat in expression file. **total**: indicator variable for the presence (1) or absence (0) of a significant difference in total expression. **cis**: indicator variable for the presence (1) or absence (0) of a signficant difference in cis-regulation. **goOutDegree**: number of GO-SLIM terms associated with the gene. **goTERMS**: names of GO-SLIM terms associated with the gene.

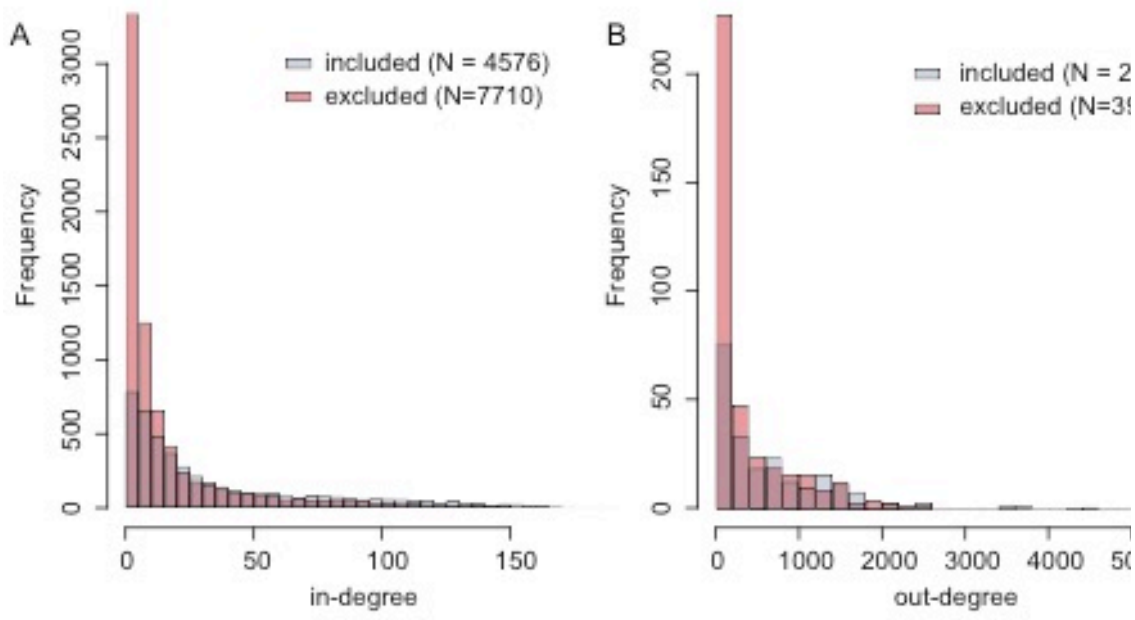

Figure 2.9. Genes excluded from the regulatory network tend to have low in-degree and low out-degree. (**A**) Overlapping histograms comparing the distributions of in-degree values for genes included (blue) and excluded (magenta) from the Marbach et al. (2012) network by the merging process described in Figure 2.8 are shown. (**B**) Overlapping histograms comparing the distributions of out-degree values for genes included (blue) and excluded (magenta) from the Marbach et al. (2012) network by the merging process described in Figure 2.8 are shown.

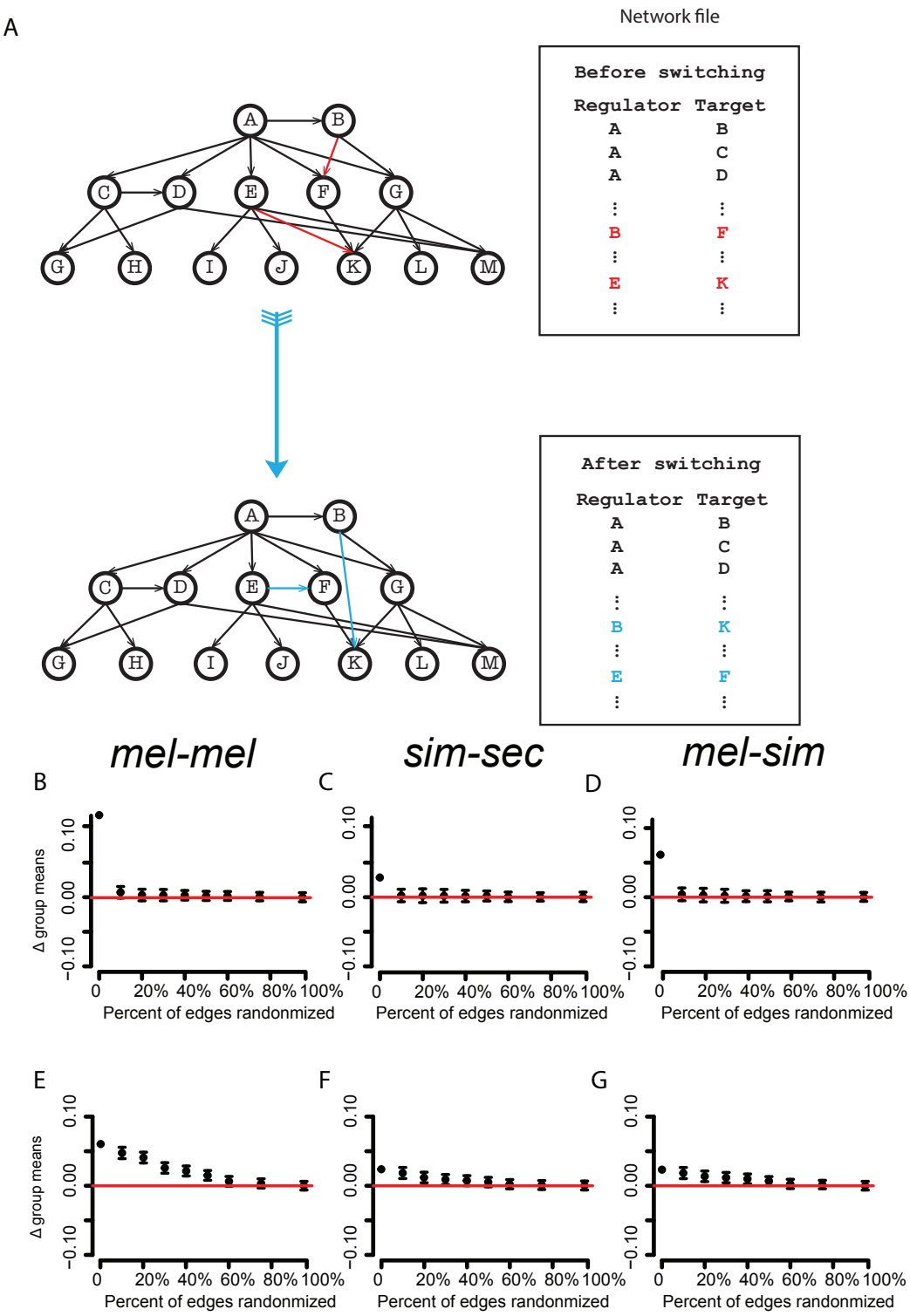

Figure 2.10. Network randomization. To assess the sensitivity of the results shown in Figure 2.1 to errors in the network structure, we switched 10%, 20%, 30%, 40%,

50%, 60%, 75%, or 95% of the edges in the network following the degree-preserving network randomization "algorithm A" described in Milo et al (2002) and depicted in (**A**). Although we find this to be an intuitive way to test the sensitivity of these results to errors in the network structure and this edge switching study has been used in prior work (Milo et al. 2002, Milo et al. 2003, Iorio et al. 2016), the properties comparing null models generated in this way to patterns of gene expression (which include covariance among genes from the true network structure) have not been established and these results should therefore be interpreted with this caveat in mind (Churchill and Doerge 2008). To shuffle edges in the network while keeping the in-degree and out-degree associated with each gene constant, we randomly picked two edges in the network data file and swapped the target genes between the two edges as long as neither of the two new edges created already appeared in the network. This procedure is illustrated in the figure, with a pair of randomly selected edges shown in red in both the datafile and the corresponding network as well as the change in these connections after swapping target genes shown in blue in both the datafile and corresponding network below. This procedure was repeated until the desired number of edges (e.g., 10%, 20%, etc) was altered. Edges were sampled without replacement so that a given edge could only be shuffled once. The script used to perform this shuffling is provided as Scripts/NETSHUFFLE.py, which can be run with Scripts/COMMANDS.sh. Note that this is a computationally intensive script that required a full day to run on the author's personal computer. (**B-D**) After switching edges in the regulatory network, the difference in the proportion of target genes with expression differences between transcription factors with and without expression differences was calculated. This randomization was repeated 200 times for each percentage of switched edges tested. The mean difference in this proportion (Y-axis) is shown for each percentage of network connections randomized (X-axis) for the *mel-mel* (**B**), *sim-sec* (**C**), and *mel-sim* (**D**) comparisons. Error bars on each point indicate two standard deviations around the mean for each set of 200 permutations. We observed that switching as few as 10% of the edges in the network eliminated the significant difference in the proportion of target genes with significant expression differences between transcription factors with and without significant expression differences. (**E-G**) Using the same set of networks with switched edges as in **B-D**, we calculated the difference in the proportion of regulators (transcription factors) with expression differences between target genes with and without expression differences. The mean difference in this proportion (Y-axis) is shown for each percentage of network connections randomized (X-axis) for the *mel-mel* (**E**), *sim-sec* (**F**), and *mel-sim* (**G**) comparisons. Error bars on each point indicate two standard deviations around the mean for each set of 200 permutations. The differences in the proportion of transcription factors with significant expression differences between target genes with and without expression differences were reduced significantly after switching

30% or more of the edges in the network in all three comparisons. The sensitivity of these metrics to changes in the network topology suggests that the *D. melanogaster* transcriptional regulatory network developed by Marbach et al. (2012) is reliable and largely conserved among the species studied.

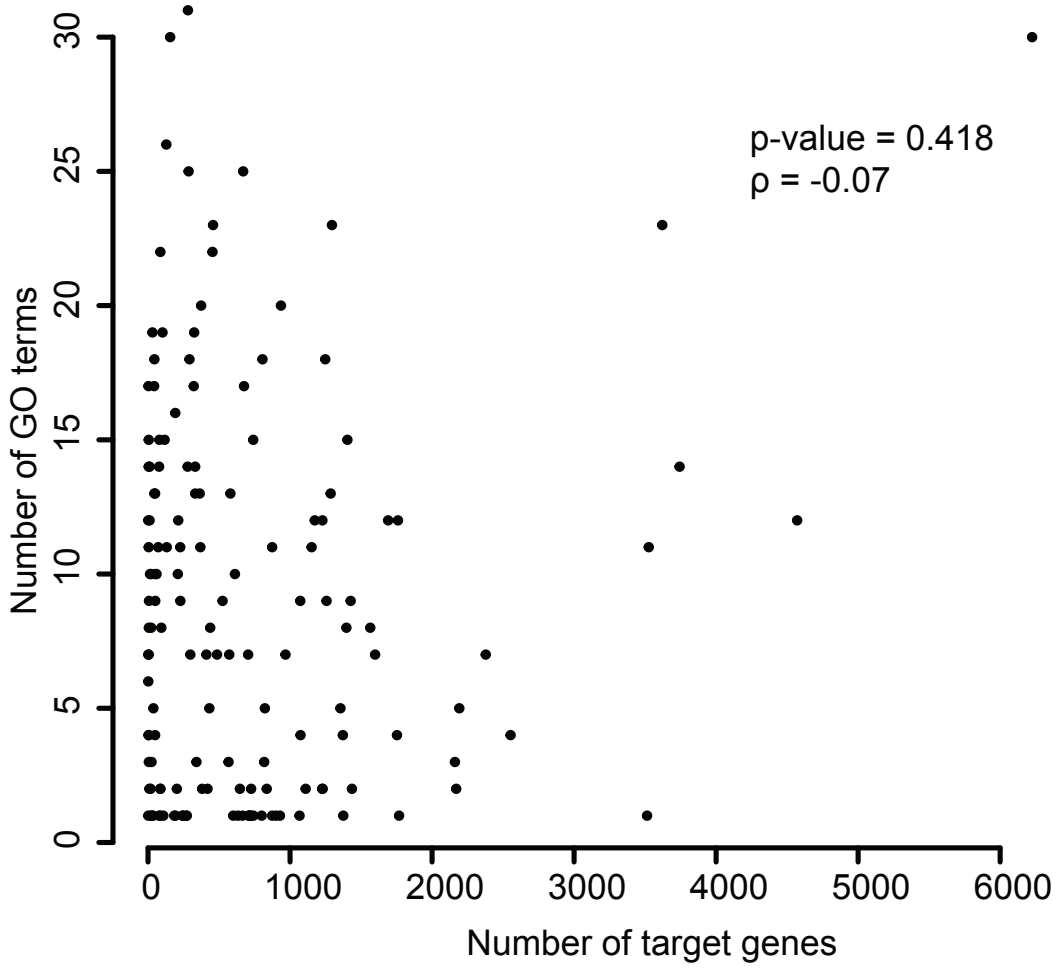

Figure 2.11. Comparing proxies for pleiotropy. The relationship between number of GO SLIM terms and number of target genes associated with transcription factors is shown. Spearman's rank correlation coefficient (ρ) and the associated p-value are also shown.

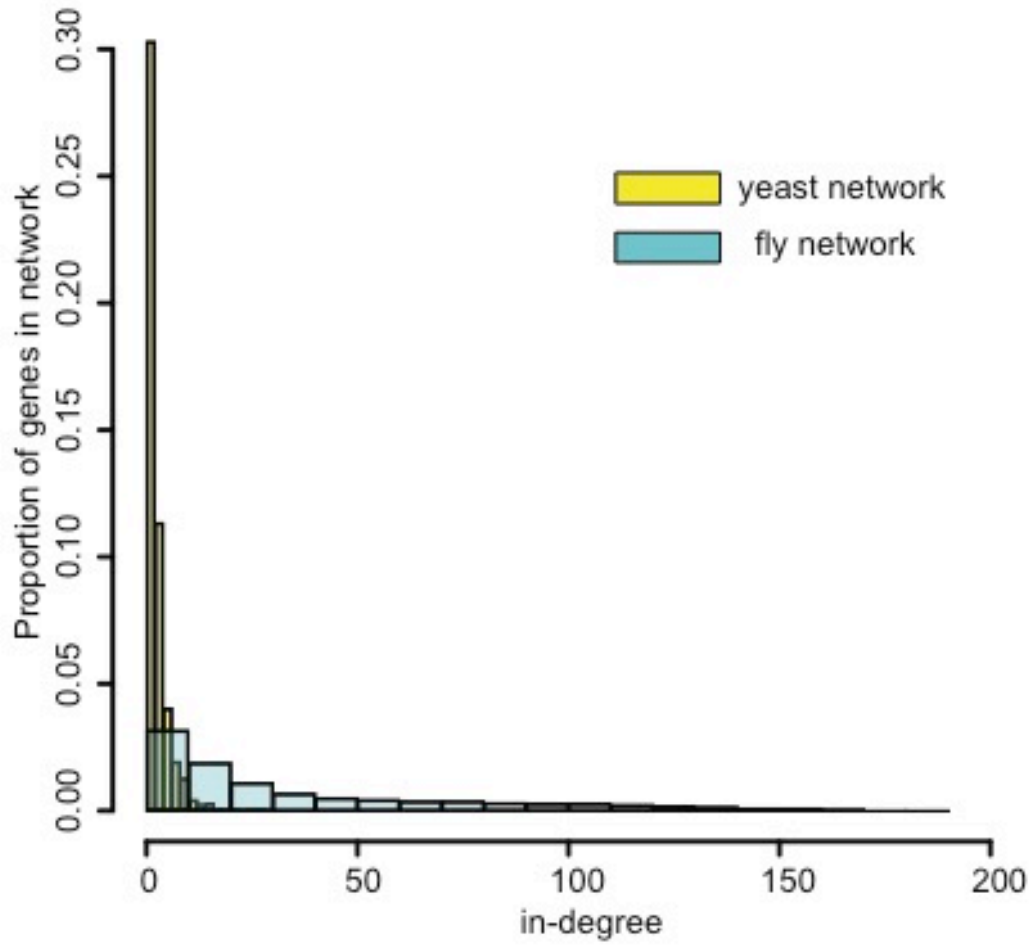

Figure 2.12. In-degree distributions differ between the transcriptional regulatory networks of flies and yeast. Histograms summarizing distributions of in-degree from *Saccharomyces cerevisiae* (yellow) and *Drosophila melanogaster* (blue) regulatory networks are shown

Chapter III

Constructing the yeast transcriptional regulatory network and examine its role in the evolution of gene expression in related yeast species

Abstract

A transcriptional regulatory network is composed of regulatory interactions between transcription factors and their target genes, which are important for the regulation of gene expression. A full description of the structure of the transcriptional regulatory network could benefit our understanding on various aspects of gene regulation. In this study, following a previously developed method, we built a regulatory network using multiple data sources from *Saccharomyces cerevisiae*. Compared to previously published networks, our inferred network contains more regulatory interactions, and achieves better performance when we used previously developed metrics to examine the quality of the network. We then used our inferred network to study whether the connective properties of regulatory network are associated with differences in gene expression across diverged *Saccharomyces* species. Specifically, we compared the in-degree (number of regulators for a gene) and the out-degree (number of targets for a transcription factor) to the differences in gene expression between two strains of *S. cerevisiae,* between *S. cerevisiae* and *S. paradoxus,* between *S. cerevisiae* and *S. mikatae,* between *S. cerevisiae* and *S. bayanus*. We found that increasing in-degree was associated with increasing differences in both the expression level (mRNA abundance) and the *cis*-regulation for the

comparison between two strains of *S. cerevisiae*, but had no statistically significant relationship with either quantity in the three between-species comparisons. We also found that out-degree had no statistically significant relationship with differences in neither the expression level nor *cis*-regulation. The conclusion for in-degree is not consistent with our previous study, in which we found that increasing in-degree was associated with decreasing differences in both the expression level and the *cis*-regulation among diverged *Drosophila* species (*D. melanogaster, D. sechellia, D. simulans*). This inconsistency might suggest that how the number of connections influences the evolution of the expression level of a gene is affected by factors unique to the species under consideration.

Introduction

In the recent decades, it has been recognized that the evolution of gene expression plays an important role in the phenotypic evolution (Wray 2007; Stern and Orgogozo 2008; Carroll 2008). Those findings motivate the efforts to examine how the regulation of gene expression evolves over time. Transcription, which is the first step of gene expression, is regulated by interactions between transcription factors and the corresponding binding sites (or *cis* elements). Collections of regulatory interactions are often represented by the transcriptional regulatory network (Zhu et al. 2007). The structural properties of the transcriptional regulatory network have been predicted to impact the evolution of the gene expression (Promislow 2005; Borneman et al. 2006; Landry et al. 2007; Macneil and Walhout 2011). It is thus interesting to examine whether those properties could

influence the observed pattern of the evolution of gene expression using empirical datasets.

One important connective property is the number of regulators for a gene (in-degree). Two studies have examined whether and how in-degree is associated to the divergence in the regulation of gene expression (Kopp and McIntyre 2012; Yang and Wittkopp). Yang and Wittkopp (2017) used datasets generated from related *Drosophila* species (*D. melanogaster, D. sechellia, D. simulans*) and found that increasing number of regulators was associated with decreasing differences in both gene expression and the *cis*-regulation. This conclusion is consistent with the hypothesis that the expression levels of genes with more regulators are on average more stable, due to the fact that the effect of interrupting a single regulatory connection could be alleviated by the coordinated changes in other regulators (Macneil and Walhout 2011). This hypothesis is supported by the discovery that master regulators in development have more regulators, and the expression levels of those master regulators are more stable compared to other genes (Borneman et al. 2006; Batada and Hurst 2007). In the other study, Kopp and Mclntyre (2012) observed an opposite trend using datasets from two *Saccharomyces* species (*S. cerevisiae* and *S. paradoxus*). They found that increasing in-degree was associated with increasing differences in *cis*-regulation. Their conclusion is consistent with the hypothesis based on mutational target size, which predicts that the expression levels of genes with more regulators are more likely to change over time, since a random mutation has higher chance to hit one of the regulators and result in changes in expression. This hypothesis is supported by several studies in yeast (Promislow 2005; Landry et al. 2007)

showing that the sensitivity of the expression level to spontaneous mutations is positively correlated with the estimated number of *trans*-regulatory factors.

The inconsistency of conclusions between yeast and fly might reflect true biological differences between the two groups of species. First, the average number of regulators of a gene in one of the yeast transcriptional regulatory networks used in Kopp and Mclntyre is smaller than the same quantity in fly transcriptional regulatory network (Yang and Wittkopp. 2017). This fact is consistent with the observation that unicellular organisms have smaller intergenic regions, and thus might harbor less *cis* regulatory elements (Nelson and Hersh 2004; Kristiansson et al. 2009; Suga et al. 2013). Also, multicellular organisms have more complicated organizations of the regulatory network to accommodate the needs for more complicated regulation on gene expression across development stages and different tissues (Davidson and Erwin 2006). Since the potential robustness conferred by the presence of multiple transcription factors depends on the number of regulators, it is possible that in-degree is not sufficiently high in yeast so that changes in gene expression caused by genetic changes in a single regulator could not be buffered due to the limited number of other regulators available to compensate. Instead, mutational target size becomes the dominant force in determining the pattern of the divergence in gene expression..

However, we could not exclude the possibility that the inconsistency in conclusions for in-degree between yeast and fly is a consequence of the differences in how the regulatory networks utilized in the two studies were constructed. Compared to the fly network used in Yang and Wittkopp (2017), the regulatory networks used in Kopp and Mclntyre (2012) have one drawback. Only one type of dataset was used to infer the

network for all networks analyzed in Kopp and Mclntyre (2012). As suggested by other studies (reviewed in Marbach, Costello, et al. 2012), incorporating multiple different types of data sources can increase the accuracy of the predicted regulatory interactions, and this point was examined in Marbach et al. (2012). In addition, more recent datasets that are useful for inferring the regulatory network in *Saccharomyces* species have been generated after Kopp and Mclntyre (2012). Thus, to better compare relationships between in-degree and differences in gene expression in multiple species, it is more reasonable to use the same network inference method to construct the regulatory network in different species.

In both Yang and Wittkopp. (2017) and Kopp and Mclntyre. (2012), it was shown that the number of targets (out-degree) for a transcription factor had no significant relationship with neither differences in gene expression nor differences in *cis*-regulation. Those conclusions are not consistent with the predictions from the hypothesis of pleiotropy, which suggests that the expression level of the transcription factors modulating more traits are more stable over evolutionary time. Changes in the expression level of those transcription factors might have more deleterious effects, since more morphological or biochemical characteristics might be affected. We also expect to test whether this conclusion still holds with an updated regulatory network in yeast.

In this study, we inferred a transcriptional regulatory network for S. cerevisiae using the method described in Marbach et al. (2012). We collected similar types of datasets used for building the fly network, and combined them with the statistical method in Marbach 2012. We examined the quality of the inferred regulatory network with two metrics used by Marbach (2012), and we found that our inferred regulatory network has

better performance compared to existing yeast networks. Then, we combined our newly inferred network with a dataset describing differences in gene expression/cis-regulation between pairs of multiple Saccharomyces species (S. cerevisiae, S. paradoxus, S. mikatae and S. bayanus) (Metzger 2017). The statistical pipeline to generate the expression dataset is similar to Coolon et al. (2014), which is the expression dataset used in Yang and Wittkopp. (2017). We examined whether and how in-degree and out-degree were associated with differences in gene expression and cis-regulation. Our results suggested that in-degree has a complicated pattern of association with the evolution in gene expression over time, which demonstrates that.

Results

Constructing a transcriptional regulatory network using multiple sources of data from S.cerevisiae

A high quality network including the majority of regulators and target genes is a prerequisite for examining the impact of regulatory interactions on gene expression evolution. To generate such a network for analysis of *Saccharomyces* species, we first inferred a transcriptional regulatory network from datasets collected in *Saccharomyces cerevisiae,* using the methods developed by Marbach et al. (2012). In Marbach et al. (2012), the authors integrated both physical and functional regulatory interactions datasets to reconstruct the regulatory network in *Drosophila melanogaster*. We collected similar types of public available datasets in *Saccharomyces cerevisiae*. The physical interactions datasets consist of (1) evolutionary conserved binding motif instances for 156 transcription factors across 7 *Saccharomyces* species (Daily et al. 2011), and (2)

whole-genome binding profiles for 106 transcription factors defined by ChIP-array or ChIP-seq experiments (Venters et al. 2011; Lickwar et al. 2012; Carrillo et al. 2012; Cai et al. 2013). The functional interactions datasets consist of (1) gene expression profiles across multiple environments or stress conditions (Gasch et al. 2000; Kvitek et al. 2008; Skelly et al. 2013), and (2) whole-genome profiles for 20 histone modification markers across several environments (Kurdistani et al. 2004; Pokholok et al. 2005). In addition, we collected information on transcriptional response upon deletion of 263 transcription factors from a whole-genome gene knockout dataset (Hu et al. 2007).

Before integrating different datasets to infer regulatory network, we first built feature specific regulatory network for each dataset. In each feature specific regulatory network, we assigned a weight to each edge for future data integration. A binary value was assigned to each edge in ChIP related feature specific network, which represents whether there exists a ChIP peak identified for the TF within 500 bp around transcription start site (TSS) of the target gene. The weights of motif feature specific network were from evolutionary conservation score calculated in Daily 2011 (Daily et al. 2011). Weights for gene expression and histone modification feature specific networks were calculated as correlation scores between TF and target gene expression profiles or histone modification profiles, respectively (see Methods). Hu et al constructed a functional transcriptional regulatory network by using the gene knockout dataset (Hu et al. 2007). Their network was directly used as the feature specific network for the knockout dataset.

We then followed methods from Marbach et al. (2012) to integrate all feature specific networks to build the whole-genome transcriptional regulatory network. We used both supervised and unsupervised statistical methods to perform the integration.

The goal is to generate a weight for each edge that can be found in any of the feature specific networks, and the weight could be considered as a confidence score representing to what extend each edge is supported by information from all datasets used. In the unsupervised method, the weight of an edge in the inferred network was calculated as the average of all weights of that edge in feature specific networks. In the supervised method, we trained a logistic regression classifier by a high-quality small-scale regulatory network constructed previously (Ma et al. 2014). We then calculated weights using the trained classifier for each edge present in any of the feature specific networks. The resulting supervised network contained 176 transcription factors, 5650 target genes and ~460,00 edges, while the resulted unsupervised network contained 151 transcription factors, 5080 target genes and ~460,00 edges.

The inferred regulatory network recovers interactions in pre-existing regulatory networks

To assess whether the supervised and unsupervised networks we inferred contain informative interactions between TFs and target genes, we first checked whether they recover interactions in previously constructed networks. We picked the three regulatory networks used in Kopp and Mclntyre (2012) and compared our two inferred networks to them. In the following discussions, the three networks will be named networkB (Balaji et al. 2006), networkL (Luscombe et al. 2004) and networkJ (Jothi et al. 2009) for convenience of discussing the results. We first counted how many edges (connections between a specific TF and a specific target gene) are shared between our inferred regulatory network and these networks. For the supervised network, 8129 edges were

shared with networkB, or 71% of the 11449 edges in networkB; for networkJ, 8427 edges were shared with network J, or 68% of the 12393 edges in networkJ; and 4187 edges were shared with networkL, or 65% of the 6441 edges in networkL. For the unsupervised network, 7327 edges were shared with networkB, or 64% for networkB; 7312 edges were shared with networkJ, or 59% in networkJ; and 3929 edges shared with networkL, or 61% in networkL. Compared to the supervised network, the unsupervised network is less effective in recovering interactions in pre-existing regulatory networks. Still, our unsupervised regulatory network can recover over 60% of the edges in the 3 pre-existing networks.

We noticed that the numbers of edges shared between two networks were limited by the total number of edges in the smaller network. To reduce the potential bias, we used network overlap enrichment score from Marbach et al. (2012) (Marbach, Roy, et al. 2012), in which number of shared edges between two networks is assumed to follow a hypergeometric distribution (see Methods). This enrichment score takes into account the size of the smaller network within each comparison, and can be used to perform statistical significant test on whether the number of observed common edges are significantly higher than expected number of shared edges using hypergeometric distribution. We calculated the overlap enrichment scores for all six comparisons (2 inferred networks (supervised and unsupervised) x 3 previously inferred networks). For the supervised regulatory network, the enrichment scores were 30.66 (networkB, p-value=4.2×10^{-29}), 23.67 (networkJ, p-value=5.4×10^{-25}) and 22.54 (networkL, p-value=7.3×10^{-19}), respectively. For the unsupervised regulatory network, the enrichment scores were 21.22 (networkB, p-value=2.5×10^{-16}), 19.37 (networkJ, p-value=2.9×10^{-13})

and 18.92 (networkL, p-value=$1.7x10^{-14}$). The highly significant enrichment scores for both the supervised network and unsupervised network suggest that interactions previously inferred by other approaches are largely recovered in our networks constructed by integrating multiple data sources.

Co-regulated genes in inferred regulatory network show functional enrichment

To further check whether our inferred regulatory networks capture biologically relevant regulatory interactions, we made use of another metric developed in Marbach 2012. The rationale behind this metric is that if two genes share more regulators, then they are more likely to participate in similar biological processes. Following this logic, we first searched for all pairs of "co-regulated" genes, with the property that the number of shared regulators are more than half of the number of all unique regulators from both genes. We then calculated whether the co-regulated genes share more similar functions than randomly picked gene pairs by enrichment analysis (see Methods). The training network shows the highest enrichment of functional similarities across co-regulated gene pairs (Figure 1, enrichment-score=2.5, p-value<0.01), supporting the use of this network in training our inferred networks. Both supervised and unsupervised networks show significant enrichment of functional similarities in co-regulated gene pairs (Figure 3.1. Supervised network: enrichment-score=1.8, pvalue<0.01; Unsupervised network: enrichment-score=1.6, pvalue<0.01). Although their enrichment scores are lower than the training network, our two inferred networks contain 9 times more interactions than the training network (~5000 edges in training network; ~45,000 in supervised and unsupervised networks). We also did the same analysis on all the existing yeast

networks. We found that our inferred networks both had higher enrichment scores than all other existing networks (Figure 3.1). The above results suggest that our inferred regulatory networks contain biologically relevant interactions.

Examining quality of the inferred supervised regulatory network in multiple Saccharomyces species.

Next, we examined whether evolution of gene expression could be affected by number of connections in a transcriptional regulatory network. Since in both validation metrics presented above, supervised network performed better than unsupervised network, all following results were generated using supervised network. Statistically significant gene expression differences within and between related *Saccharomyces* species were taken from Metzger et al. (2017). In the study, Metzger et al used the RNA-seq data collected within (Schaefke et al. 2013) and between (Schraiber et al. 2013) species to calculate gene expression differences in four comparisons: BY4741 and RM11 strains of *S.cerevisiae* (*cer-cer*), *S.cerevisiae* and *S.paradoxus* (*cer-par*), *S.cerevisiae* and *S.mikatae* (*cer-mik*), *S.cerevisiae* and *S.bayanus* (*cer-bay*). In the same study, *cis*-regulatory differences were also calculated for all four comparisons, which provide a more direct readout of relationship between transcription factors and target genes. All the following analysis were restricted to 3034 genes out of 5652 genes in the supervised regulatory network and have both expression differences and *cis*-regulatory differences across all four comparisons in Metzger et al. (2017). Among the 3034 genes included, 64 are transcription factors, and all 3034 genes have at least one regulator. Details about how

the information from regulatory network and gene expression differences dataset were merged could be found in Figure 3.10.

Since the regulatory network was constructed using data from *S. cerevisiae*, we first checked whether this network could be used in all four yeast species. Using a method from Yang and Wittkopp. (2017), we examined whether it is appropriate to use the same regulatory network in multiple species (Figure 3.2). The rationale for this validation method is described as follows. If connections within a transcriptional regulatory network capture functional relevant interactions between TFs and target genes, then we expect that the target genes of transcription factors with changed expression are more likely to also change their expression. We found that transcription factors with changed expression in within species comparisons (*cer-cer*) had on average a greater proportion of target genes also change expression (Figure 3.2A). Following a similar argument, on the other hand, the regulators of genes with changed expression are more likely to also change expression. We also found support for this expectation for the comparison between two strains of *S. cerevisiae* (Figure 3.2B). These results suggest that our supervised regulatory network captures functional informative regulatory interactions between transcription factors and target genes.

To examine whether our network could be used on species other than *S. cerevisiae*, we redid the above analyses using the expression data from the three other yeast species (*S. paradoxus, S. mikatae* and *S. bayanus*). For all four comparisons, we observed the same pattern as in *S. cerevisiae* (Figure 3.3). However, we did observe that differences in mean proportion of targets/regulators that have expression differences between group of transcription factors/genes with and without expression differences decrease with

increasing divergence time (Figure 3.9). The implication of this observation will be discussed in Discussion section.

In-degree correlates with differences in gene expression and cis-regulatory differences within species, but not between species in yeast

We next examined whether increasing in-degree was associated with decreasing differences in gene expression, as suggested from our previous studies using *Drosophila* data. First, we compared the in-degree distributions between genes with ($N_{cer\text{-}cer}$ = 1030, $N_{cer\text{-}par}$ = 2359, $N_{cer\text{-}mik}$ = 2364, $N_{cer\text{-}bay}$ = 2453) and without ($N_{cer\text{-}cer}$ = 2004, $N_{cer\text{-}par}$ = 675, $N_{cer\text{-}mik}$ = 670, $N_{cer\text{-}bay}$ = 581) statistically significant expression differences in each of the four comparisons (Figure 3.4). Unlike what we observed in *Drosophila* study, median of in-degree distributions is higher for group of genes with expression differences in *cer-cer* within species comparison (Figure 3.4A, Wilcoxon rank sum test, $P_{cer\text{-}cer} = 1.5 \times 10^{-26}$). However, in all between species comparisons, differences in medians between the two group of genes were not statistically significant (Figure 3.4B-D, Wilcoxon rank sum test, $P_{cer\text{-}par} = 0.53$, $P_{cer\text{-}mik} = 0.058$, $P_{cer\text{-}bay} = 0.26$).

To directly understand whether in-degree correlates with gene expression differences, we examined how proportion of genes with statistically significant expression differences changed with in-degree. Consistent with what we observed above, increasing in-degree was found to be associated with increasing of proportion of genes with statistically significant expression difference in the within species comparison (Logistic regression, $P_{cer\text{-}cer} < 2 \times 10^{-16}$, $\beta = 0.06$), while not significant relationships were found for all between species comparisons (Logistic regression, $P_{cer\text{-}par} = 0.46$, $P_{cer\text{-}mik}$ =

0.21, $P_{cer\text{-}bay} = 0.14$). We also compared in-degree to absolute magnitude of gene expression differences. For within species comparison, magnitude of gene expression differences showed a statistically significant positive correlation with in-degree using Spearman non-parametric correlation test (Figure 3.4E, $P_{cer\text{-}cer} = 4.6 \times 10^{-13}$, $\rho = 0.13$). However, no significant correlations were detected in all three between species comparisons (Figure 3.4F-H, $P_{cer\text{-}par} = 0.41$, $P_{cer\text{-}mik} = 0.38$, $P_{cer\text{-}bay} = 0.98$). The above analysis showed that for the within species comparison, in-degree had opposite relationships with gene expression differences between yeast and fly. However, we failed to detect any significant relationships between in-degree and gene expression differences for all three between species comparisons in yeast.

We repeated our analysis but using *cis*-regulatory differences instead of gene expression differences, to examine whether direct readout of expression divergence rate in *cis*-elements might correlate with in-degree. The results were consistent with what we found for gene expression differences (Figure 3.5. In within species comparison, genes with *cis*-regulatory differences ($N_{cer\text{-}cer} = 431$) had on average higher in-degree that those without ($N_{cer\text{-}cer} = 2603$) (Figure 3.5A, Wilcoxon rank sum test, $P_{cer\text{-}cer} = 4.1 \times 10^{-15}$). Also, increasing in-degree was associated with both increasing proportion of gene with *cis*-regulatory differences and magnitude of the differences (Logistic regression, $P_{cer\text{-}cer} < 2 \times 10^{-16}$, $\beta = 0.05$; Figure 3.5E, Spearman's correlation test, $P_{cer\text{-}cer} = 2.3 \times 10^{-5}$, $\rho = 0.07$). However, in all three between species comparisons, no significant relationship between in-degree and *cis*-regulatory differences was detected using any analysis approach (Figure 3.5B-D, Wilcoxon rank sum test, $P_{cer\text{-}par} = 0.34$, $P_{cer\text{-}mik} = 0.063$, $P_{cer\text{-}bay} = 0.53$; Logistic regression, $P_{cer\text{-}par} = 0.52$, $P_{cer\text{-}mik} = 0.041$, $P_{cer\text{-}bay} = 0.51$; Figure 3.5F-H,

Spearman's correlation test, $P_{cer\text{-}par} = 0.77$, $P_{cer\text{-}mik} = 0.35$, $P_{cer\text{-}bay} = 0.87$). The consistency between conclusions from both expression differences and *cis*-regulatory differences suggested that effects of in-degree on the evolution of *cis*-regulatory activity are at least partially responsible for the observed relationship between in-degree and differences in gene expression (Yang and Wittkopp 2017)

Out-degree does not correlate with either expression differences or cis-regulatory differences in both within and between species comparisons.

Next, we examined whether out-degree had a significant association with gene expression evolution, as predicted by the pleiotropy theory. First, we compared median out-degree between transcription factors with ($N_{cer\text{-}cer} = 20$, $N_{cer\text{-}par} = 48$, $N_{cer\text{-}mik} = 43$, $N_{cer\text{-}bay} = 51$) and without ($N_{cer\text{-}cer} = 44$, $N_{cer\text{-}par} = 16$, $N_{cer\text{-}mik} = 21$, $N_{cer\text{-}bay} = 13$) expression differences (Figure 3.6A-D). Consistent with what we observed in *Drosophila* species, median out-degrees were not statistically significant different between the two group of transcription factors in all four comparisons (Figure 3.6A-D, Wilcoxon rank sum test, $P_{cer\text{-}cer} = 0.27$, $P_{cer\text{-}par} = 0.45$, $P_{cer\text{-}mik} = 0.81$, $P_{cer\text{-}bay} = 0.93$). We then examined whether out-degree was associated with either proportion of transcription factors with expression differences or magnitude of expression differences. Both analysis suggested that out-degree did not provide useful information on observed pattern of gene expression differences over multiple evolutionary time points (Logistic regression, $P_{cer\text{-}cer} = 0.89$, $P_{cer\text{-}par} = 0.55$, $P_{cer\text{-}mik} = 0.60$, $P_{cer\text{-}bay} = 0.74$; Figure 3.6E-H, Spearman's correlation test, $P_{cer\text{-}cer} = 0.40$, $P_{cer\text{-}par} = 0.35$, $P_{cer\text{-}mik} = 0.93$, $P_{cer\text{-}bay} = 0.67$).

We repeated all the analysis on out-degree described above, but used *cis*-regulatory differences instead of gene expression differences. Consistent with what we observed for expression differences, median out-degrees were not statistically significant different between transcription factors with ($N_{cer\text{-}cer}$ = 6, $N_{cer\text{-}par}$ = 41, $N_{cer\text{-}mik}$ = 48, $N_{cer\text{-}bay}$ = 49) and without ($N_{cer\text{-}cer}$ = 58, $N_{cer\text{-}par}$ = 23, $N_{cer\text{-}mik}$ = 16, $N_{cer\text{-}bay}$ = 15) *cis*-regulatory differences (Figure 3.7A-D, Wilcoxon rank sum test, $P_{cer\text{-}cer}$ = 0.05, $P_{cer\text{-}par}$ = 0.63, $P_{cer\text{-}mik}$ = 0.55, $P_{cer\text{-}bay}$ = 0.57). We then examined whether out-degree was associated with either proportion of transcription factors with *cis*-regulatory differences or magnitude of *cis*-regulatory differences. Neither analysis suggested that out-degree was associated with *cis*-regulatory differences (Logistic regression, $P_{cer\text{-}cer}$ = 0.30, $P_{cer\text{-}par}$ = 0.81, $P_{cer\text{-}mik}$ = 0.44, $P_{cer\text{-}bay}$ = 0.84; Figure 3.7E-H, Spearman's correlation test, $P_{cer\text{-}cer}$ = 0.66, $P_{cer\text{-}par}$ = 0.34, $P_{cer\text{-}mik}$ = 0.86, $P_{cer\text{-}bay}$ = 0.35). Taken together, similar to what we observed in *Drosophila* species, out-degree did not have a statistically significant relationship with either gene expression differences or *cis*-regulatory differences.

Discussion

The regulatory interactions between transcription factors and *cis* regulatory elements play a critical role in transcriptional regulation. In the current study, we inferred a transcriptional regulatory network from *S. cerevisiae*, and utilized it to answer the question that whether the connective properties of genes within a regulatory network could be associated with the evolution of gene expression. We found that increasing number of regulators (in-degree) was associated with increasing differences in the expression level / *cis* regulation in the comparison between two strains of *S. cerevisiae*.

However, the relationship between in-degree and differences in gene expression was not significant in all three comparisons between two diverged *Saccharomyces* species (*S. cerevisiae* and *S. paradoxus, S.cerevisiae* and *S. mikatae, S. cerevisiae* and *S. bayanus*). In addition, we found that the number of targets (out-degree) for a transcription factor did not show significant association with the differences in gene expression. The conclusion for in-degree is inconsistent with the result from our previous study in *Drosophila* species (Yang and Wittkopp. 2017), in which we found that increasing in-degree was associated with decreasing differences in both gene expression level and *cis* regulation. However, our result for in-degree is similar to the Kopp and Mclntyre. (2012), in which they found that in-degree was positively correlated with *cis*-regulation between *S. cerevisiae* and *S. paradoxus*. Below, we will discuss the implications by comparing results from all those studies.

Network in-degree does not have a consistent relationship with patterns of evolution of gene expression

The fact that transcription of a gene is regulated by multiple transcription factors might either suppresses or enhances the evolution of gene expression. One hypothesis is that the expression levels of genes with more regulators are more sensitive to new mutations, since there exist more sites in the genome that can change expression. Our result from the comparison between two strains of *S. cerevisiae,* as well as study from Kopp and Mclntyre. (2012), are both consistent with this hypothesis. However, as we pointed out in Yang and Wittkopp. (2017), the changes in expression for a single transcription factor do not necessarily result in changes in expression for target gene. Functional redundancy

among regulators are widely observed in different biological systems (Wu and Lai 2015; Kuntz et al. 2012; Macneil and Walhout 2011). From this perspective, if a gene has multiple transcription factors, changes in expression caused by genetic disruptions of a single regulator could be "compensated" by other regulators through coordinated changes in the expression level or binding strength. This assumption leads to the hypothesis that genes with more regulators are more robust in the evolution of gene expression, which is consistent with our findings in flies (Yang and Wittkopp. 2017). It is thus important to understand why the conclusions are not consistent in different groups of related species.

One possible explanation is that the average numbers of regulatory input for a gene are different between yeast and fly. By using similar network construction approach and the statistical significance criteria, we found that distributions of the in-degree were different between fly and yeast (Figure 3.8). Genes in the fly regulatory network have on average more regulators than genes in the yeast regulatory network. The implication from this observation is that there are fewer numbers of replacement transcription factors available to compensate for genetic perturbations in transcriptional regulatory input program in yeast. In addition, because the effective population size of yeast population in wild is much larger than flies, mutations could accumulate faster in yeast populations. In combination, the higher number of available genetic variations and less robustness to new mutations of gene expression lead to the observed positive association between network in-degree and differences in gene expression in within species comparison in yeast.

In our previous analysis using fly datasets, conclusions on relationship between the in-degree and differences in gene expression and/or *cis*-regulation are consistent over evolutionary time. However, this is not true for the current analysis in yeast. We

provided two potential explanations for this observation.

First, the structural organization of the regulatory network might experience changes across evolutionary time. Although it is found that the core regulatory programs are conserved between related species (Erwin and Davidson 2009; Rebeiz et al. 2015), the strength of the regulatory interactions might change over time. For example, although a regulatory interaction might exist in all related species, the regulator might have a changed level of impact on determining the expression level of the target gene. We illustrated this idea in Figure 3.9, where we plotted the differences in average proportion of target/regulators with changed expression between regulators/target genes with and without expression differences against sequence divergence of each comparison, using data from Figure 3.2. Although the differences across all comparisons are larger than zero, the magnitudes decrease for comparisons involving pairs of species with longer divergence time. This observation suggests that although on average the network in *S. cerevisiae* might still provide useful information for other yeast species, the strength of interactions might change, the fact of which results in the decreased relatedness between the expression level of transcription factors and target genes. It is thus possible that relationship between in-degree and gene expression evolution might not be detectable because of the changes in interaction strength.

Second, although genetic perturbations in any regulator could result in changes in the expression level of the target gene, the levels of effects do not have to be equal for different regulators. More specifically, it is possible that only a few transcription factors for a gene, when acquire genetic changes, could result in detectable changes in the expression level of the target gene. Thus, whether or not a gene acquires diverged

expression might completely depend on the evolutionary states of transcription factors that have stronger impact on its expression level. In another word, mutations in different regulators or even in different regions of one regulator could have very different effect sizes on target gene's expression level. The evolutionary fate of gene expression is thus under the control of at least two forces: number of mutations that have the potential to change expression and effect size of each mutation. In longer divergence time, since more mutations can be accumulated, there are higher chances that larger effect genetic variants are hit in diverged species, the fact of which obscures the detection of importance of mutational target size on gene expression evolution. From this argument, the loss of significant relationship between in-degree and gene expression evolution in comparisons involving diverged *Saccharomyces* species might be due to the imbalanced effect size of random mutations in different regulators. The complication of our conclusions in different divergence time suggests that knowing effect size or identity of mutations is as important as knowing mutational target size.

Token together, although we do not know the accurate biological reasons to explain the inconsistency in our observations, our results suggest whether and how in-degree is associated gene expression evolution depend on the specific organization and natural history of the underlying regulatory network in the biological system used.

Network out-degree appears to have minimal effect on gene expression evolution

Although the conclusions on the in-degree differ in different context, there is a good consistency for the conclusions on the out-degree in both fly and yeast analysis. Interestingly, the absence of association between out-degree and the differences in gene

expression and/or cis-regulation in all between-species comparisons is not consistent with the predictions from the hypothesis of pleiotropy. However, our results do not provide refute to pleiotropy neither. One explanation is that the out-degree is not a good estimate of pleiotropy, although this possibility is difficult to examine because there is not a consensus on what is the right measure of pleiotropy.

It should be pointed out that the number of transcription factors are limited in both studies (127 in fly analysis and 64 in yeast analysis). This might limit our power to detect a weak but significant association between the out-degree and the evolution of gene expression. In addition, it has been illustrated in many case studies that phenotypic innovation could occur through genetic changes in transcription factors (Wagner and Lynch 2008; Rebeiz et al. 2015). Thus, the evolution of gene expression for transcription factors might not be as restricted as previously expected from hypothesis of pleiotropy. Thus, the stabilizing selection on the expression level of transcription factors suggested by the hypothesis of pleiotropy might not be the only evolutionary force that shape the pattern of the divergence in the expression observed in natural populations. A better understanding of evolutionary fate of each transcription factor requires the knowledge on the involved biological functions, as well as different constraints imposed on higher-order phenotypes faced by diverged species.

Concluding remarks

Although a full description of the structural properties of the transcriptional regulatory network might provide useful information in understanding the evolution of gene expression, they are not the driving force in shaping the evolutionary pattern. More

precisely, the transcriptional regulatory network provides constraints on the potential evolutionary trajectory, while the evolutionary fate of gene expression is largely determined by other biological forces, including generation of genetic variations, selection and drift, and other population parameters. Our observations that relationships between the structural properties of the regulatory network and the evolution of gene expression are not consistent across different groups of species suggest that the biological forces that connect the regulatory network to the evolution of gene expression might be different in different systems. We propose two future directions to further dissect how the regulatory network could influence the evolution of gene expression. First, it is necessary to keep collecting genomics data on both regulatory interactions and gene expression in more species, so that we could get much more accurate descriptions on the network structure and the evolution of the expression levels. Analysis on those genomics data could generate how various aspects of the structural organization of the regulatory network are associated with evolutionary patterns. To empirically dissect what biological forces could shape those associations detected from genomic studies, experimental systems in which both structure of the regulatory network and evolutionary forces can be controlled should be used to better understand the functional basis of the observed patterns. Combination of those two approaches will produce more informative insights on understanding of the evolution of gene expression in a network context.

Materials and Methods

Gene annotations

Gene annotations are obtained from SGD (2016-11-05). Dubious ORFs or ORFs with no annotated functions have been removed from all analysis. The list of transcription factors is compiled from Table 1 in T.R.Hughes et al. (2013) (Hughes and de Boer 2013).

Input datasets for regulatory network inference

Physical binding datasets were obtained from two major sources: *Saccharomyces* Genome Database (http://www.yeastgenome.org/) and Venters et al. (2011) (Venters et al. 2011). Interaction datasets from SGD was downloaded from yeastmine (http://yeastmine.yeastgenome.org/) website. The downloaded data file contained regulatory interactions estimated from various techniques and experimental growth conditions. All regulatory interactions from ChIP-array or ChIP-seq experiments were retained for further analysis.

Motif dataset was extracted from Daily et al. (2011) (Daily et al. 2011). The authors first used Positional Weight Matrix (PWM) to look for all possible motif instances in the yeast genome. Then they used whole genome alignment of 7 yeast species to estimate the evolutionary conservation level of each motif instance, with the rationale that conserved motif instances were more likely to be functional binding sites. Conservation scores calculated for each motif instance were extracted from the website MotifMap (http://motifmap.ics.uci.edu/). For cases where multiple motifs of the same transcription factor were identified for a target gene, only the motif with highest conservation score was retained for further analysis.

Gene expression datasets were obtained from Skelly et al. (2013), Kvitek et al. (2008) and Gasch et al. (2000). All datasets were processed and provided on SGD website (http://www.yeastgenome.org/).

Whole genome histone modification marker datasets were obtained from Kurdistani et al. (2004) and Pokholok et al. (2005). In Kurdistani dataset, 11 histone acetylation profiles were collected in YEPD environment. In Pokholok dataset, 8 histone methylation and acetylation profiles were collected. Processed datasets from both studies can be downloaded from SGD website ((http://www.yeastgenome.org/).

Dataset on transcriptome response to gene deletions was collected from Hu et al. (2007). The functional regulatory network constructed in the study was directly used as one of the feature specific network.

Constructing feature specific regulatory network

Before constructing each feature specific regulatory network, all possible interactions between transcription factors (TF) and genes were recorded. If a TF-gene pair was not contained in one the feature specific regulatory network, a default value of 0.05 was assigned to the pair in the corresponding feature specific network, as described in Marbach et al. (2012) (Marbach, Roy, et al. 2012). Each feature specific network associated each TF-gene pair with a weight, representing how strong the interaction was supported from each dataset. In summary, two types of physical binding networks were created: ChIP network and motif network; three types of functional regulatory network were created: co-expression network, co-modification network and knockout network.

For ChIP network, if a binding peak of a transcription factor is within 500 bp upstream or downstream of transcription start site (TSS) of a gene, a value of 1 was given to the TF-gene pair. Peak information for interactions downloaded from SGD was extracted from associated studies.

For motif network, the conservation scores calculated from Daily et al. (2011) were used for weights associated with interactions. We followed advice from Marbach et al. (2012). Specifically, if no motif instance of a TF was found within 500 bp upstream or downstream of TSS of a gene, a value of 0.0 was used as the weight for the TF-gene pair. In all other cases, conservation scores were increased by 0.1, with the restriction that weights can be no larger than 1.0.

For co-expression networks, one feature specific network was constructed for each of the three datasets used. Pearson correlations of processed gene expression levels were calculated for all cases in which both the transcription factor and target gene were assayed in the expression dataset, and were used as weights associated with interactions.

For co-modification networks, correlation of histone modification was calculated for each TF-gene pair if data existed for both. Specifically, genomic sequence of a gene was divided into five regions: 1kb upstream of 5' UTR, 5' UTR, gene body, 3' UTR and 1kb downstream of 3' UTR. For each histone modification marker, a binary value 1 was given a region if significant modification signal was detected. For Pokholok dataset, the significance level was arbitrarily determined as >1.2 of relative enrichment score calculated in the study, which is used in analysis in the original paper.

For knockout network, the functional regulatory network from Hu et al. (2007) was directly used to construct feature specific network. Specifically, if interaction of a TF-

gene pair was supported in Hu network, a binary value of 1 was then given to that pair in the knockout network.

Integrating feature specific regulatory networks

Feature specific regulatory networks were integrated by two approaches: unsupervised learning method and supervised learning method. Before integration, TF-gene pairs with support from none of the feature specific regulatory networks were removed.

In unsupervised method, weights for interactions in the resulted network were calculated by average of all weights in feature specific regulatory networks. Edges with top 4% of weights were retained to build the final unsupervised network, so that number of edges in unsupervised network was comparable to supervised network.

In supervised method, a logistic classifier was trained to calculate weights of interactions. The training set used in this study was from Ma et al. (2014). Specifically, the network generated in Ma 2014 contained 4034 edges out of 102 x 2060 possible interactions (after filtering dubious ORFs). Thus, the training set contained 4034 true positives and ~200000 negatives. We followed advice from Marbach et al. (2012), and used 10-fold cross validation with stratified and balanced learning approaches to train the logistic classifier. The model was trained using LogisticRegressionCV function in sklearn.linear_model python module (http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegressionCV.html, python version 3.4.2). The final weights were calculated as average weights from 500 iterations of model training. Edges with weight >0.6 were retained for further analysis, in order to be consistent with Marbach et al. (2012).

Network overlap enrichment scores

Significance of overlaps between two networks was tested using hypergeometric distribution, described as follows. First, genes not in the smaller network were removed from analysis for both networks. We then recorded (1) n, number of interactions shared by the two networks; (2) N, number of all possible interactions between group of TFs and group of genes (equals a x b, where a is the number of all TFs and b is the number of all genes); (3) A, number of interactions in the larger network; (4) B, number of interactions in the smaller network. Number of common edges between two networks should follow a hypergeometric distribution, and the expected number is m = A*B/N. Overlap enrichment score was calculated as m / n, and significance of enrichment was tested using the underlying hypergeometric distribution.

Go term enrichment scores

In GO term enrichment, all co-regulated gene pairs were first identified from the network under interest. Two genes were considered co-regulated if >50% of their unique regulators were common. For each co-regulated gene pair, proportion of common Gene Ontology terms (number of common GO terms / total number of unique GO terms for both genes) was calculated. To remove the redundancy in annotations, GO terms from SGD database were curated using previously published methods to retain enough specificity in distinguishing different functional categories (Myers et al. 2006). Proportions of common GO terms for all co-regulated gene pairs were collected. Same analysis was repeated for 100 randomized network. In each randomization, names were

shuffled for TFs and genes separately, while the connections in network were retained. Enrichment score for GO term similarity of co-regulated genes was calculated as ratio of average proportions of common GO terms between true observation and randomization. Significance of the enrichment was tested using Wilcoxon rank sum test between true observations and randomizations.

Comparing gene expression and cis-regulatory activity among strains and species

Differences in mRNA transcript abundance ("gene expression") and relative *cis*-regulatory activity between the BY4741 and RM11 strains of *S.cerevisiae* (*cer-cer*), *S.cerevisiae* (YHL068) and *S.paradoxus* (CBS432), *S.cerevisiae* (YHL068) and *S.mikatae* (IFO1815), *S.cerevisiae* (YHL068) and *S.bayanus* (CBS7001) were taken from the analysis of in Metzger et al. (2017). These data include comparisons of gene expression as well as comparisons of relative *cis*-regulatory activity inferred by comparing relative allelic expression in F1 hybrids produced by crossing each pair of strains or species (Schaefke et al. 2013; Schraiber et al. 2013) . We analyzed the 3034 genes (including 64 transcription factors) that are retained for measuring allele-specific expression used by Metzger et al. (2017) and also appeared in the supervised regulatory network.

Comparing network properties to differences in gene expression and cis-regulation

Analyses shown in Figures 3.2, 3.3, 3.4A-D, 3.5A-D, 3.6A-D and 3.7A-D, compare the presence or absence of statistically significant (FDR = 0.01) differences in gene expression or *cis*-regulatory activity described in Metzger et al. (2017) to relationships

among genes in the network (Figure 3.2 and 3.3), in-degree of all target genes (Figures 3.4 and 3.5) and out-degree of all transcription factors (Figure 3.6 and 3.7). Non-parametric Wilcoxon rank sum tests were used to compare median in-degree and out-degree between sets of genes with and without statistically significant differences in gene expression or *cis*-regulation for each pair of strains or species examined as well as to compare the proportion of target genes with differential expression between transcription factors with and without differential expression and vice versa. These tests evaluated the null hypothesis of no association between in-degree or out-degree and differences in gene expression or *cis*-regulation. Logistic regressions were also used to compare an indicator variable representing whether or not a gene had a statistically significant difference in gene expression and/or *cis*-regulatory activity in a given comparison to its in-degree or out-degree. These tests were performed using the glm function in R with the options "family=binomial, link=logit", which uses a Z-score to assess the statistical significance of the factor being tested; a significant test indicates that the factor tested (e.g., in-degree or out-degree) has statistically significant predictive ability for which genes have significant expression differences. The null hypothesis in each case was that the factor tested was not a significant predictor of differences in expression or *cis*-regulation.

Spearman's rank correlation coefficients were used to test for a significant correlation between the $\log_2$ transformed magnitude of the differences in gene expression or *cis*-regulatory activity reported in Metzger et al. (2017) and a gene's in-degree or out-degree. The null hypothesis for this test is that there is no relationship between a gene's in-degree

or out-degree and the magnitude of its expression difference between strains or species.

Results from these tests are shown in Figures 3.4E-H, 3.5E-H, 3.6E-H, and 3.7E-H.

Statistical analyses

All statistical analyses were performed in R v3.2.2 (RCoreTeam 2016).

References

Balaji S, Babu MM, Iyer LM, Luscombe NM, Aravind L. 2006. Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. J. Mol. Biol. 360:213–227.

Batada NN, Hurst LD. 2007. Evolution of chromosome organization driven by selection for reduced gene expression noise. Nature Genetics 39:945–949.

Borneman AR, Leigh-Bell JA, Yu H, Bertone P, Gerstein M, Snyder M. 2006. Target hub proteins serve as master regulators of development in yeast. Genes Dev. 20:435–448.

Cai L, McCormick MA, Kennedy BK, Tu BP. 2013. Integration of multiple nutrient cues and regulation of lifespan by ribosomal transcription factor Ifh1. Cell Rep 4:1063–1071.

Carrillo E, Ben-Ari G, Wildenhain J, Tyers M, Grammentz D, Lee TA. 2012. Characterizing the roles of Met31 and Met32 in coordinating Met4-activated transcription in the absence of Met30. Mol. Biol. Cell 23:1928–1942.

Carroll SB. 2008. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. Cell 134:25–36.

Coolon JD, McManus CJ, Stevenson KR, Graveley BR, Wittkopp PJ. 2014. Tempo and mode of regulatory evolution in Drosophila. Genome Res. 24:797–808.

Daily K, Patel VR, Rigor P, Xie X, Baldi P. 2011. MotifMap: integrative genome-wide maps of regulatory motif sites for model species. BMC Bioinformatics 12:495.

Davidson EH, Erwin DH. 2006. Gene regulatory networks and the evolution of animal body plans. Science.

Erwin DH, Davidson EH. 2009. The evolution of hierarchical gene regulatory networks. Nature Reviews Genetics 10:141–148.

Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO. 2000. Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. Mol. Biol. Cell 11:4241–4257.

Hu Z, Killion PJ, Iyer VR. 2007. Genetic reconstruction of a functional transcriptional regulatory network. Nature Genetics 39:683–687.

Hughes TR, de Boer CG. 2013. Mapping Yeast Transcriptional Networks. Genetics 195:9–36.

Jothi R, Balaji S, Wuster A, Grochow JA, Gsponer J, Przytycka TM, Aravind L, Babu MM. 2009. Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture. Molecular Systems Biology 5:294.

Kopp A, McIntyre LM. 2012. Transcriptional network structure has little effect on the rate of regulatory evolution in yeast. Mol. Biol. Evol. 29:1899–1905.

Kristiansson E, Thorsen M, Tamás MJ. 2009. Evolutionary forces act on promoter length: identification of enriched cis-regulatory elements. Molecular biology and ….

Kuntz SG, Williams BA, Sternberg PW, Wold BJ. 2012. Transcription factor redundancy and tissue-specific regulation: evidence from functional and physical network connectivity. Genome Res. 22:1907–1919.

Kurdistani SK, Tavazoie S, Grunstein M. 2004. Mapping Global Histone Acetylation Patterns to Gene Expression. Cell 117:721–733.

Kvitek DJ, Will JL, Gasch AP. 2008. Variations in stress sensitivity and genomic expression in diverse S. cerevisiae isolates.McVean G, editor. PLoS Genet. 4:e1000223.

Landry CR, Lemos B, Rifkin SA, Dickinson WJ, Hartl DL. 2007. Genetic Properties Influencing the Evolvability of Gene Expression. Science 317:118–121.

Lickwar CR, Mueller F, Hanlon SE, McNally JG, Lieb JD. 2012. Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. Nature 484:251–255.

Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M. 2004. Genomic analysis of regulatory network dynamics reveals large topological changes. Nature 431:308–312.

Ma S, Kemmeren P, Gresham D, Statnikov A. 2014. De-Novo Learning of Genome-Scale Regulatory Networks in S. cerevisiae.la Fuente de A, editor. PLoS ONE 9:e106479.

Macneil LT, Walhout AJM. 2011. Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. Genome Res. 21:645–657.

Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Consortium TD, Kellis M, Collins JJ, et al. 2012. Wisdom of crowds for robust gene network inference. Nature Methods 9:796–804.

Marbach D, Roy S, Ay F, Meyer PE, Candeias R, Kahveci T, Bristow CA, Kellis M. 2012. Predictive regulatory models in Drosophila melanogaster by integrative inference of transcriptional networks. Genome Res. 22:1334–1349.

McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp PJ. 2010. Regulatory divergence in Drosophila revealed by mRNA-seq. Genome Res. 20:816–825.

Metzger B, Wittkopp PJ, Coolon J. 2017. Evolutionary dynamics of regulatory changes underlying gene expression divergence among Saccharomyces species. Genome Biology and Evolution 9(4): 843-854

Myers CL, Barrett DR, Hibbs MA, Huttenhower C, Troyanskaya OG. 2006. Finding function: evaluation methods for functional genomic data. BMC Genomics 2014 15:1 7:187.

Nelson CE, Hersh BM. 2004. The regulatory content of intergenic DNA shapes genome architecture. Genome ….

Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, Lee TI, Bell GW, Walker K, Rolfe PA, Herbolsheimer E, et al. 2005. Genome-wide Map of Nucleosome Acetylation and Methylation in Yeast. Cell 122:517–527.

Promislow D. 2005. A regulatory network analysis of phenotypic plasticity in yeast. The American Naturalist.

Rebeiz M, Patel NH, Hinman VF. 2015. Unraveling the Tangled Skein: The Evolution of Transcriptional Regulatory Networks in Development. http://dx.doi.org/10.1146/annurev-genom-091212-153423 16:103–131.

Schaefke B, Emerson JJ, Wang TY, Lu M. 2013. Inheritance of gene expression level and selective constraints on trans-and cis-regulatory changes in yeast. Molecular biology and ….

Schraiber JG, Mostovoy Y, Hsu TY, Brem RB. 2013. Inferring evolutionary histories of pathway regulation from transcriptional profiling data.Teichmann SA, editor. PLoS Comput Biol 9:e1003255.

Skelly DA, Merrihew GE, Riffle M, Connelly CF, Kerr EO, Johansson M, Jaschob D, Graczyk B, Shulman NJ, Wakefield J, et al. 2013. Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. Genome Res. 23:1496–1504.

Stern DL, Orgogozo V. 2008. THE LOCI OF EVOLUTION: HOW PREDICTABLE IS

GENETIC EVOLUTION? Evolution 62:2155–2177.

Suga H, Chen Z, De Mendoza A, Sebé-Pedrós A. 2013. The Capsaspora genome reveals a complex unicellular prehistory of animals. Nature.

Venters BJ, Wachi S, Mavrich TN, Andersen BE, Jena P, Sinnamon AJ, Jain P, Rolleri NS, Jiang C, Hemeryck-Walsh C, et al. 2011. A comprehensive genomic binding map of gene and chromatin regulatory proteins in Saccharomyces. Mol. Cell 41:480–492.

Wagner GP, Lynch VJ. 2008. The gene regulatory logic of transcription factor evolution. Trends in Ecology & Evolution 23:377–385.

Wittkopp PJ, Haerum BK, Clark AG. 2004. Evolutionary changes in cis and trans gene regulation. Nature.

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. Nature Reviews Genetics 8:206–216.

Wu W-S, Lai F-J. 2015. Functional redundancy of transcription factors explains why most binding targets of a transcription factor are not affected when the transcription factor is knocked out. BMC Syst Biol 9:S2.

Yang B, Wittkopp PJ. 2017. Structure of the transcriptional regulatory network correlates with regulatory divergence in Drosophila. Molecular Biology and Evolution (2017). *in press*

Zhu X, Gerstein M, Snyder M. 2007. Getting connected: analysis and principles of biological networks. Genes Dev. 21:1010–1024.

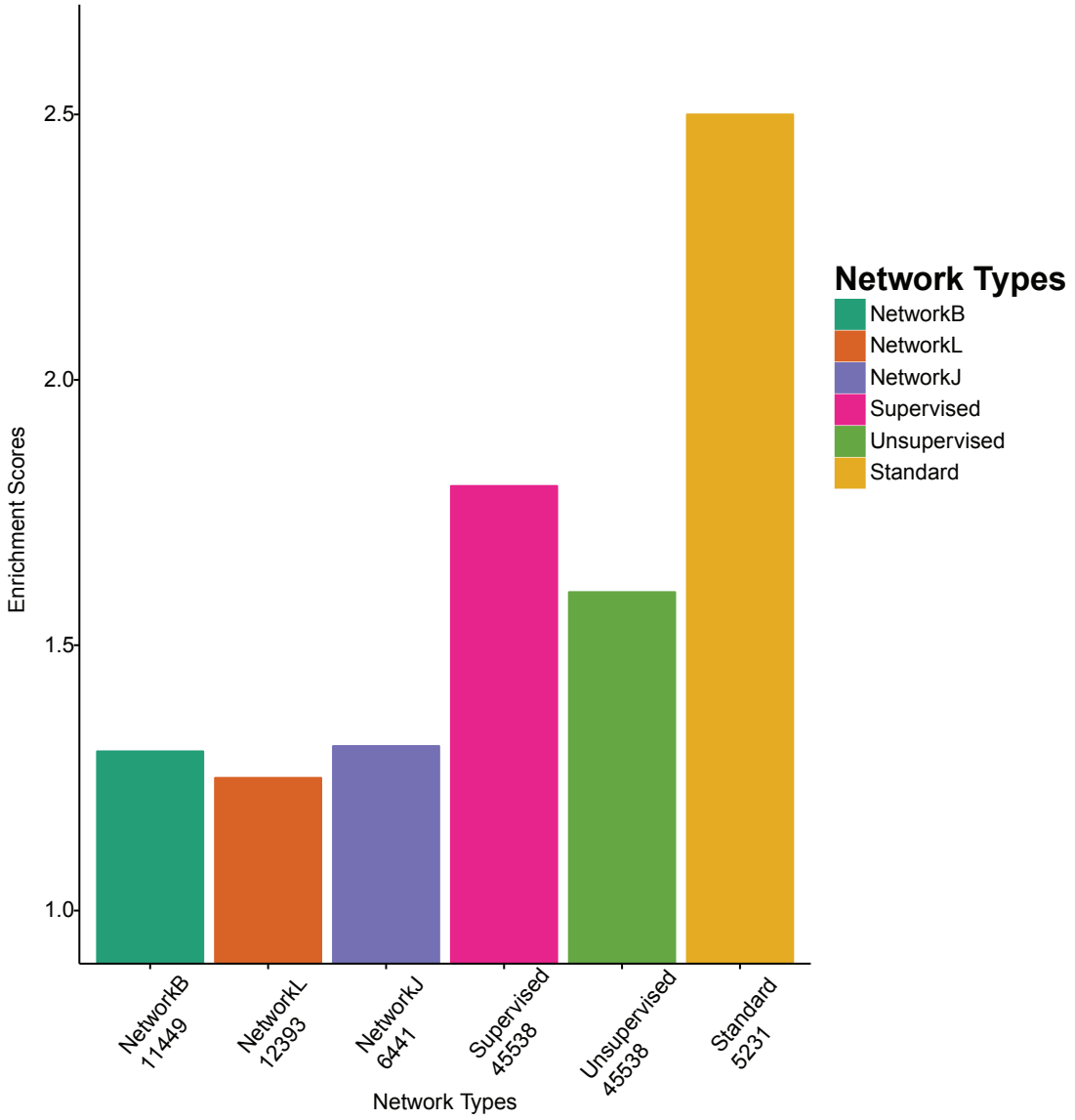

Figure 3.1. Co-regulated genes have similar biological functions in inferred regulatory network. A pair of genes is called "co-regulated" if they share more than half of their regulators. The proportion of common annotated functions (Gene Ontology terms) for each co-regulated pair was calculated, and average of this proportion across all co-regulated pairs was divided by mean of the same value from 200 randomized networks to calculate the enrichment scores shown in the figure (more details see Materials and Methods). Number of edges for each network is labeled after the network name.

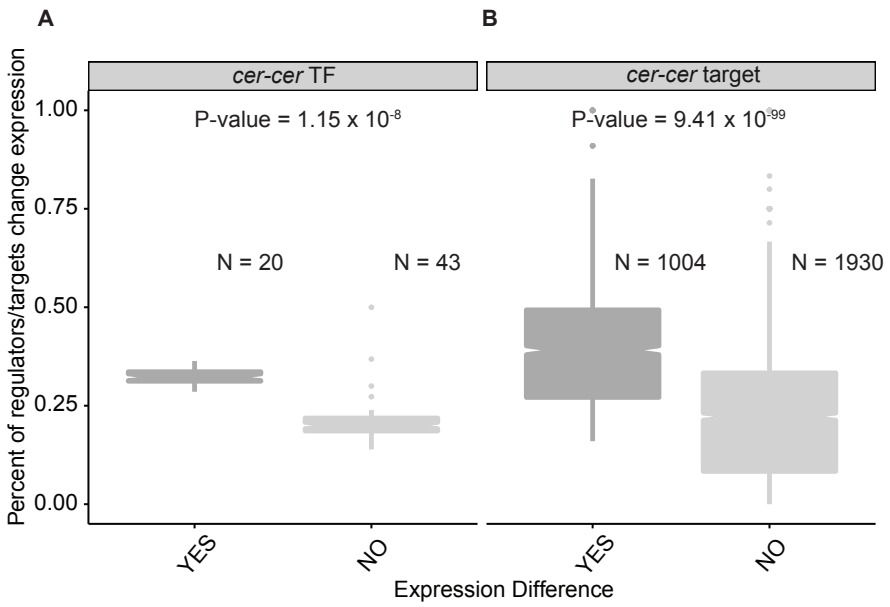

Figure 3.2. Assessment of applicability of the regulatory network in *S. cerevisiae*. (**A**) For each transcription factor (N = 64), we calculated the proportion of its target genes that showed significant expression differences between the two strains of *S. cerevisiae*. The boxplots show the distributions of these proportions for transcription factors with (dark grey) and without (light grey) significant expression differences between the two strains of *S .cerevisiae*. P-values shown are from non-parametric Wilcoxon rank sum tests, and N indicates the number of transcription factors included in each category. (**B**) For each target gene (N = 3034), we calculated the proportion of regulators (transcription factors) that showed significant expression differences between the strains or species being compared. The boxplots show the distributions of these proportions for target genes with (dark grey) and without (light grey) significant expression differences between the two strains of *S .cerevisiae.* P-values shown are from non-parametric Wilcoxon rank sum tests, and N indicates the number of target genes included in each category.

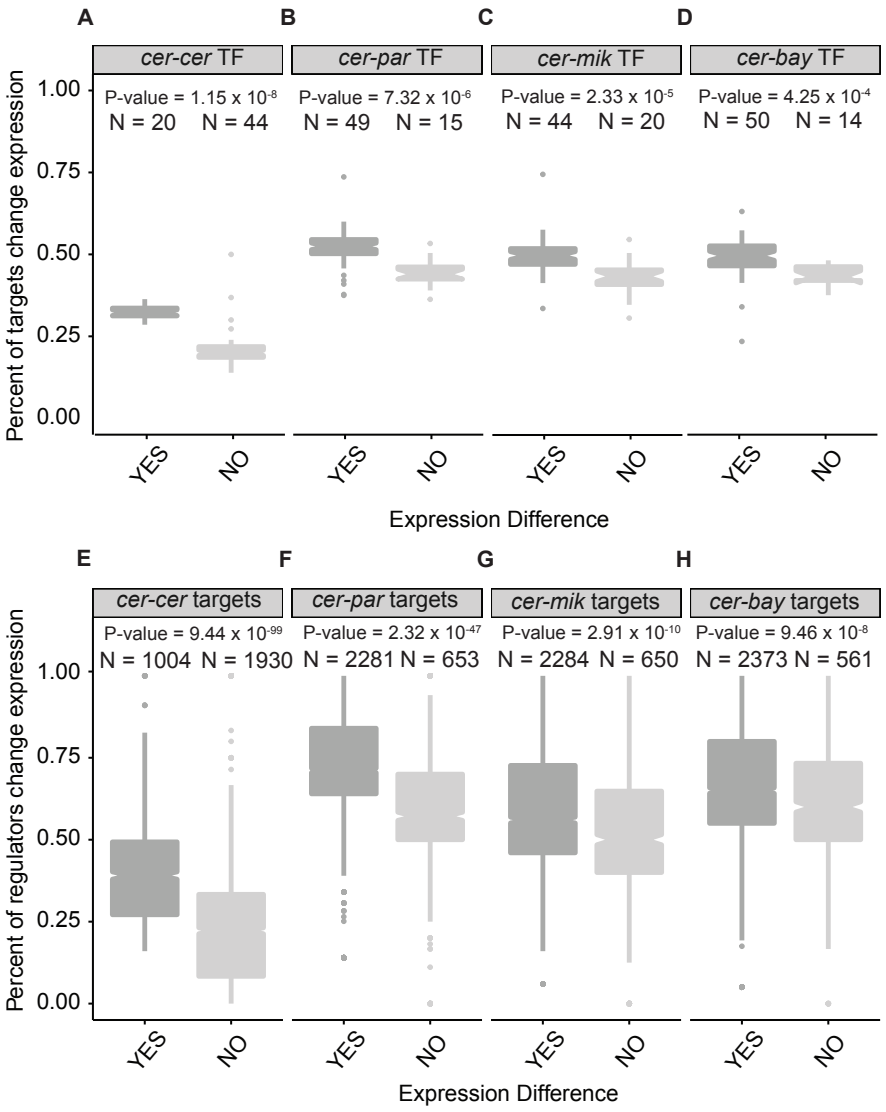

Figure 3.3. Assessment of applicability of the regulatory network in multiple *Saccharomyces* species. (**A-D**) For each transcription factor (N = 64), we calculated the proportion of its target genes that showed significant expression differences between the two strains of *S. cerevisiae*. The boxplots show the distributions of these proportions for transcription factors with (dark grey) and without (light grey) significant expression differences between the two strains of *S .cerevisiae* (**A**), between *S. cerevisiae* and *S. paradoxus* (**B**), between *S. cerevisiae* and *S. mikatae* (**C**) and between *S. cerevisiae* and *S. bayanus* (**D**). P-values shown are from non-parametric Wilcoxon rank sum tests, and N indicates the number of transcription factors included in each category. (**E-H**) For each target gene (N = 3034), we calculated the proportion of regulators (transcription factors) that showed significant expression differences between the strains or species being compared. The boxplots show the distributions of these proportions for target genes with (dark grey) and without (light grey) significant expression differences between the two

strains of *S .cerevisiae* (**E**), between *S. cerevisiae* and *S. paradoxus* (**F**), between *S. cerevisiae* and *S. mikatae* (**G**) and between *S. cerevisiae* and *S. bayanus* (**H**). P-values shown are from non-parametric Wilcoxon rank sum tests, and N indicates the number of target genes included in each category.

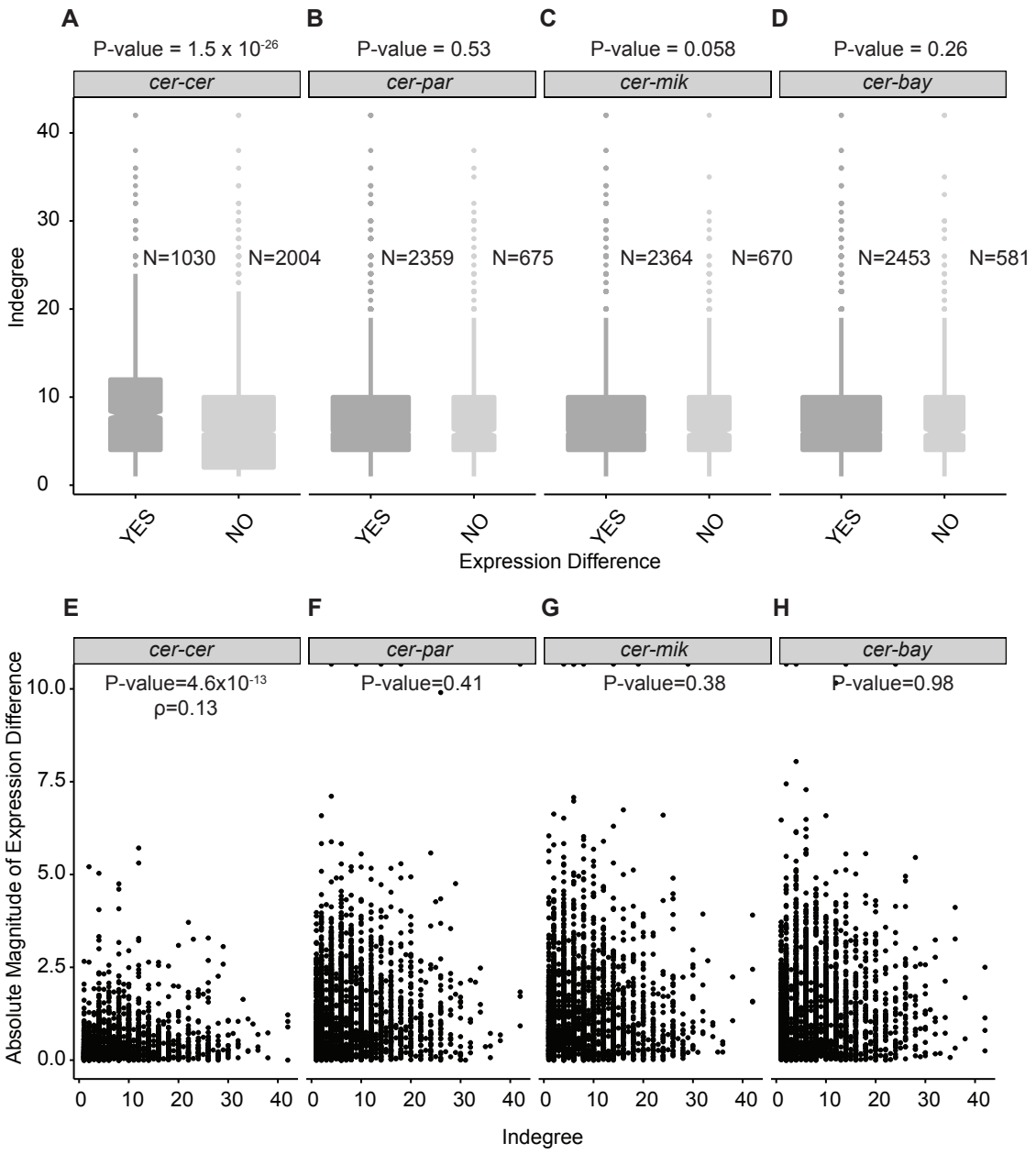

Figure 3.4. Relationship between network in-degree and differences in gene expression within species and between species. (**A-D**) Boxplots show the in-degree distributions for genes with (dark grey) and without (light grey) significant differences in gene expression in the *cer-cer* (**A**), *cer-par* (**B**), *cer-mik* (**C**) and *cer-bay* (**D**) comparisons. P-values are from non-parametric Wilcoxon rank sum tests, and N indicates the number of genes in each group. (**E-H**) Absolute magnitude of gene expression differences (Y-axis) is plotted against in-degree (X-axis) in the *cer-cer* (**E**), *cer-par* (**F**), *cer-mik* (**G**) and *cer-bay* (**H**) comparisons. Spearman's rank correlation coefficients (ρ) and associated p-values are also shown.

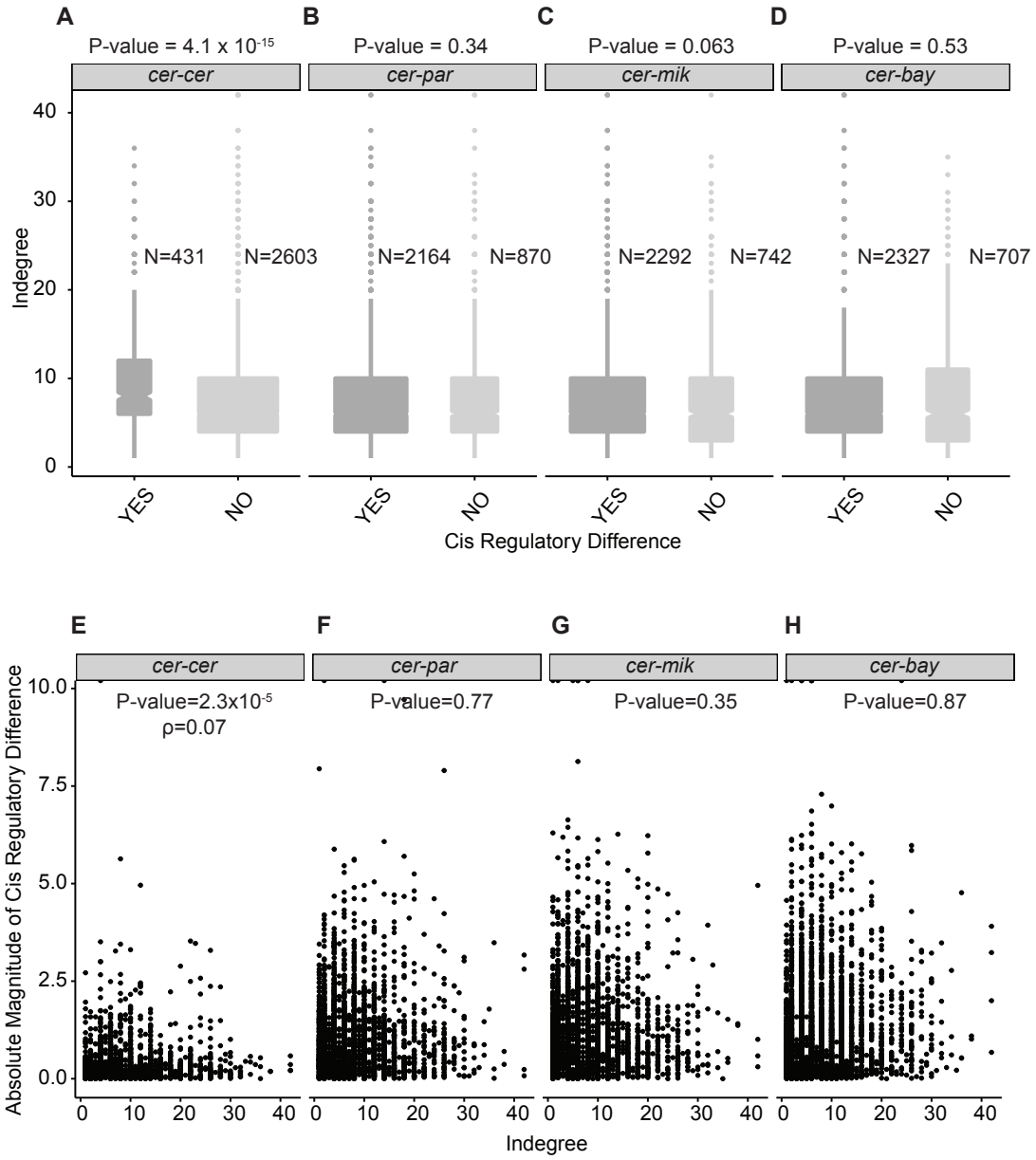

Figure 3.5. Relationship between network in-degree and differences in *cis*-regulation within species and between species. (**A-D**) Boxplots show the in-degree distributions for genes with (dark grey) and without (light grey) significant differences in *cis*-regulation in the *cer-cer* (**A**), *cer-par* (**B**), *cer-mik* (**C**) and *cer-bay* (**D**) comparisons. P-values are from non-parametric Wilcoxon rank sum tests, and N indicates the number of genes in each group. (**E-H**) Absolute magnitude of cis-regulatory differences (Y-axis) is plotted against in-degree (X-axis) in the *cer-cer* (**E**), *cer-par* (**F**), *cer-mik* (**G**) and *cer-bay* (**H**) comparisons. Spearman's rank correlation coefficients (ρ) and associated p-values are also shown.

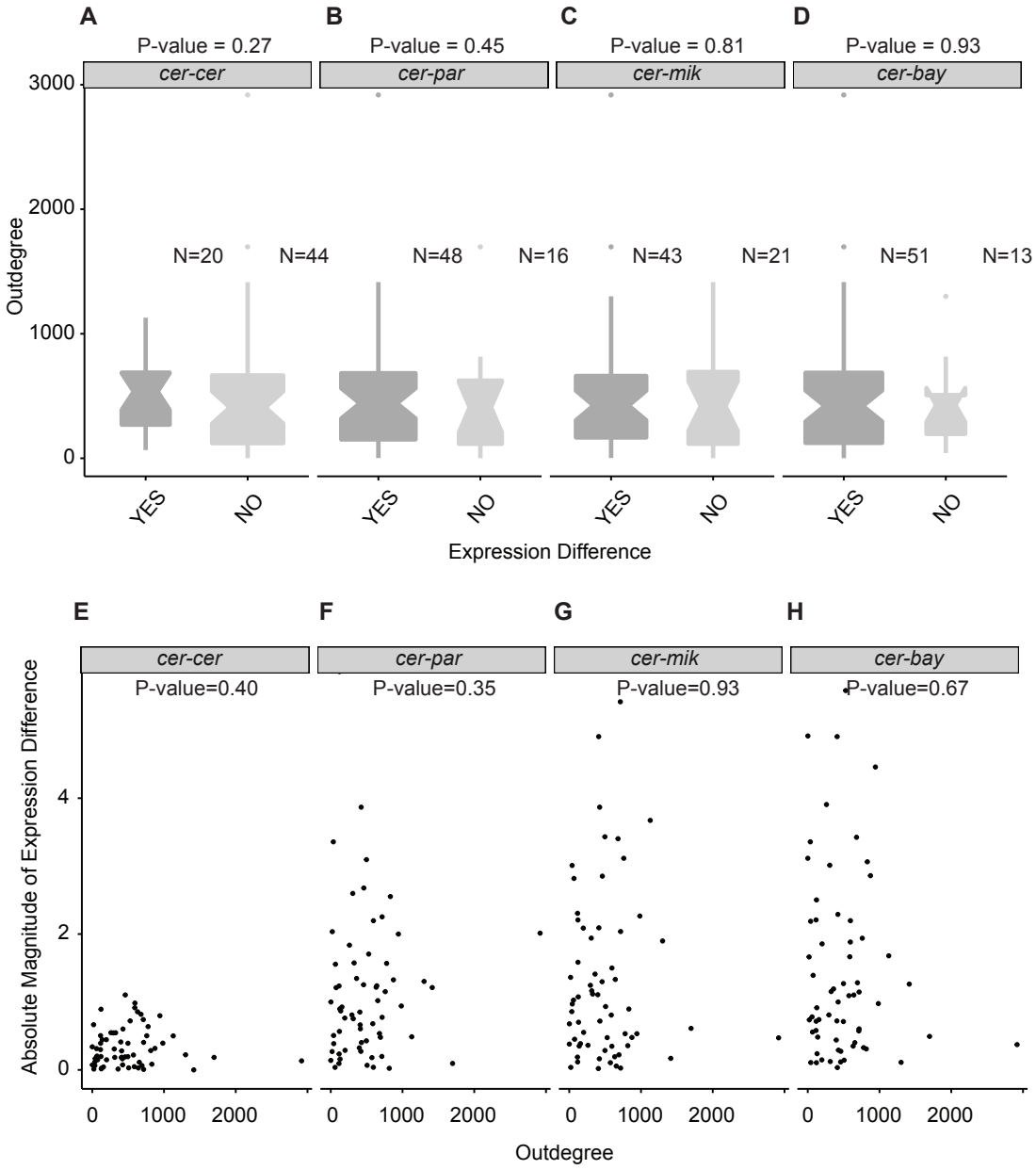

Figure 3.6. Relationship between network out-degree and differences in gene expression within species and between species. (**A-D**) Boxplots show the out-degree distributions for genes with (dark grey) and without (light grey) significant differences in gene expression in the *cer-cer* (**A**), *cer-par* (**B**), *cer-mik* (**C**) and *cer-bay* (**D**) comparisons. P-values are from non-parametric Wilcoxon rank sum tests, and N indicates the number of genes in each group. (**E-H**) Absolute magnitude of gene expression differences (Y-axis) is plotted against out-degree (X-axis) in the *cer-cer* (**E**), *cer-par* (**F**), *cer-mik* (**G**) and *cer-bay* (**H**) comparisons. Spearman's rank correlation coefficients (ρ) and associated p-values are also shown.

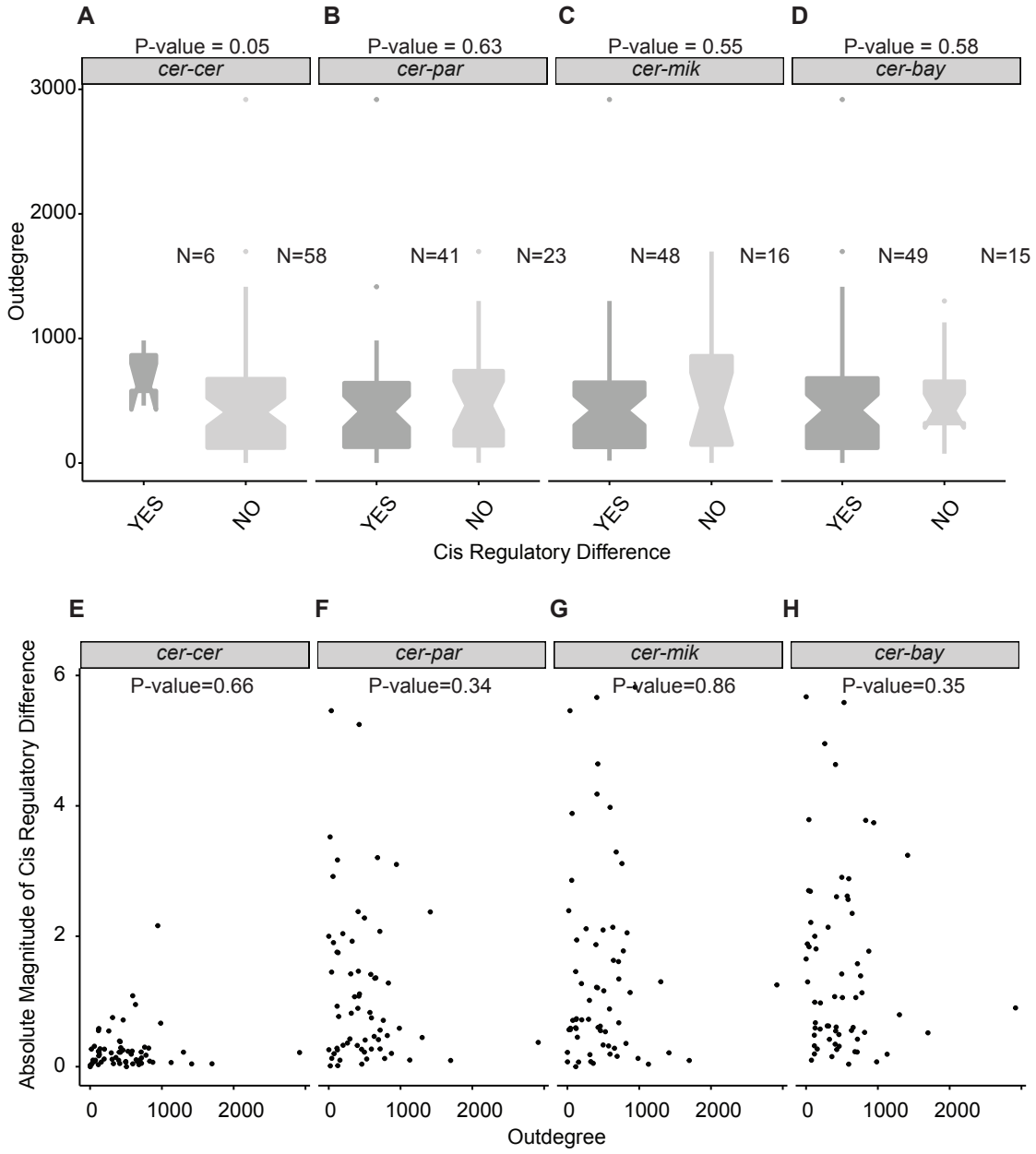

Figure 3.7. Relationship between network out-degree and differences in *cis*-regulation within species and between species. (**A-D**) Boxplots show the out-degree distributions for genes with (dark grey) and without (light grey) significant differences in *cis*-regulation in the *cer-cer* (**A**), *cer-par* (**B**), *cer-mik* (**C**) and *cer-bay* (**D**) comparisons. P-values are from non-parametric Wilcoxon rank sum tests, and N indicates the number of genes in each group. (**E-H**) Absolute magnitude of *cis*-regulatory differences (Y-axis) is plotted against out-degree (X-axis) in the *cer-cer* (**E**), *cer-par* (**F**), *cer-mik* (**G**) and *cer-bay* (**H**) comparisons. Spearman's rank correlation coefficients (ρ) and associated p-values are also shown.

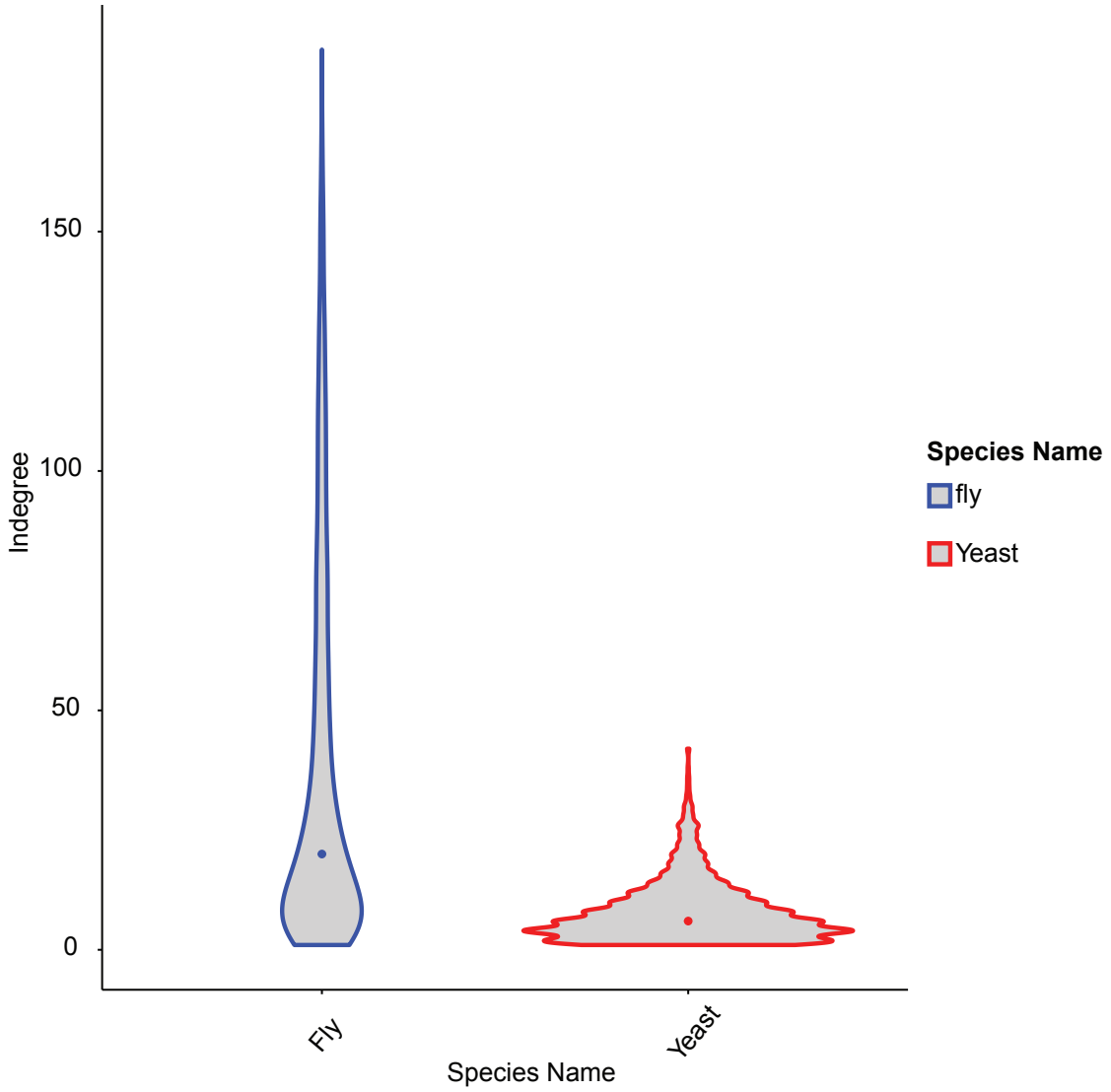

Figure 3.8. Comparing in-degree distributions between the fly and the yeast regulatory network. Violin plots showing the distribution of in-degree using either the fly network (red) and the yeast supervised network (green).

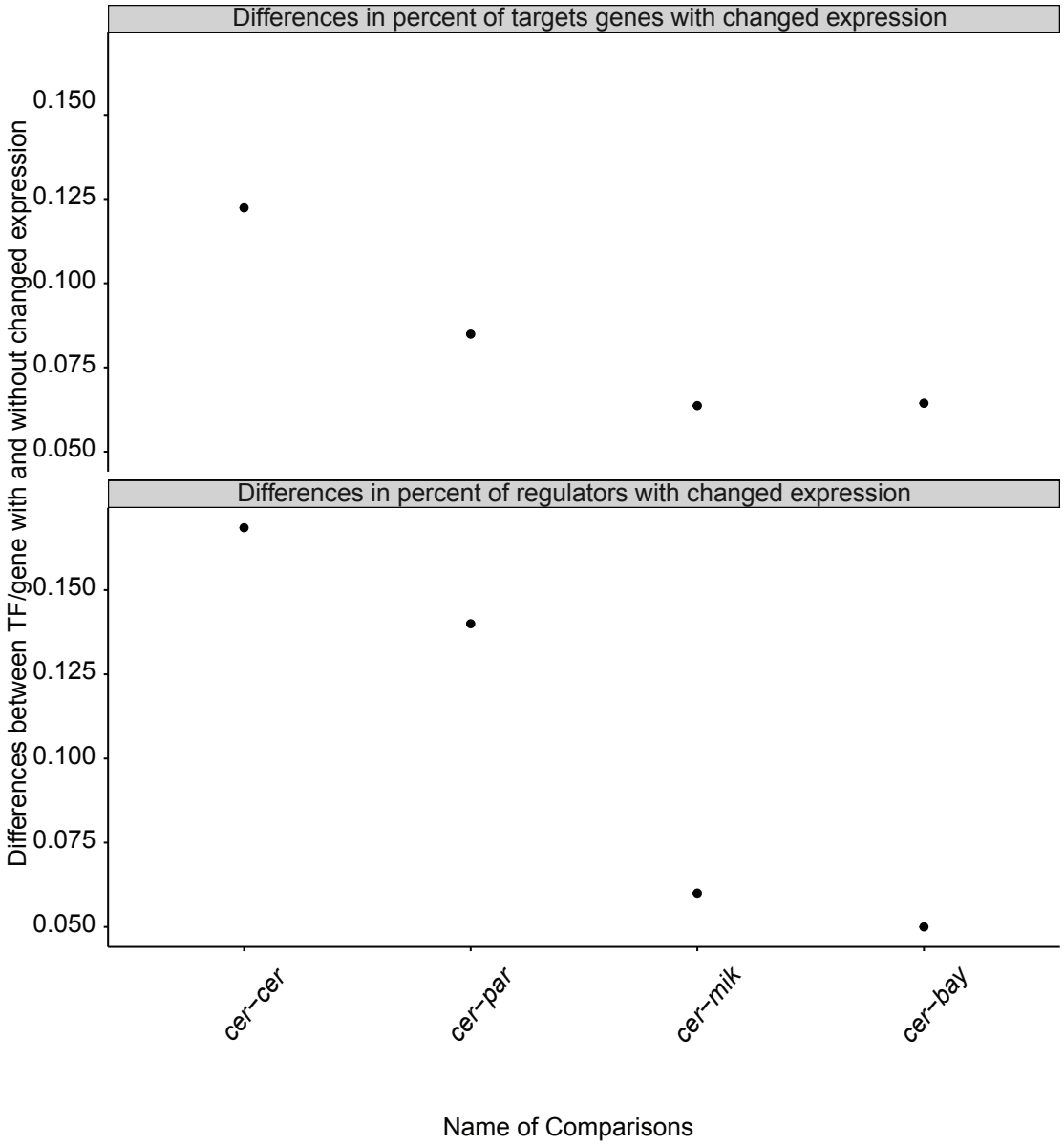

Figure 3.9. Strength of functional interactions represented in regulatory network decreased over evolutionary time. (A). Differences in average proportion of targets with changed expression between group of transcription factors with and without difference in expression were plotted for all four crosses. **(B).** Differences in average proportion of regulators with changed expression between group of target genes with and without difference in expression were plotted for all four crosses.

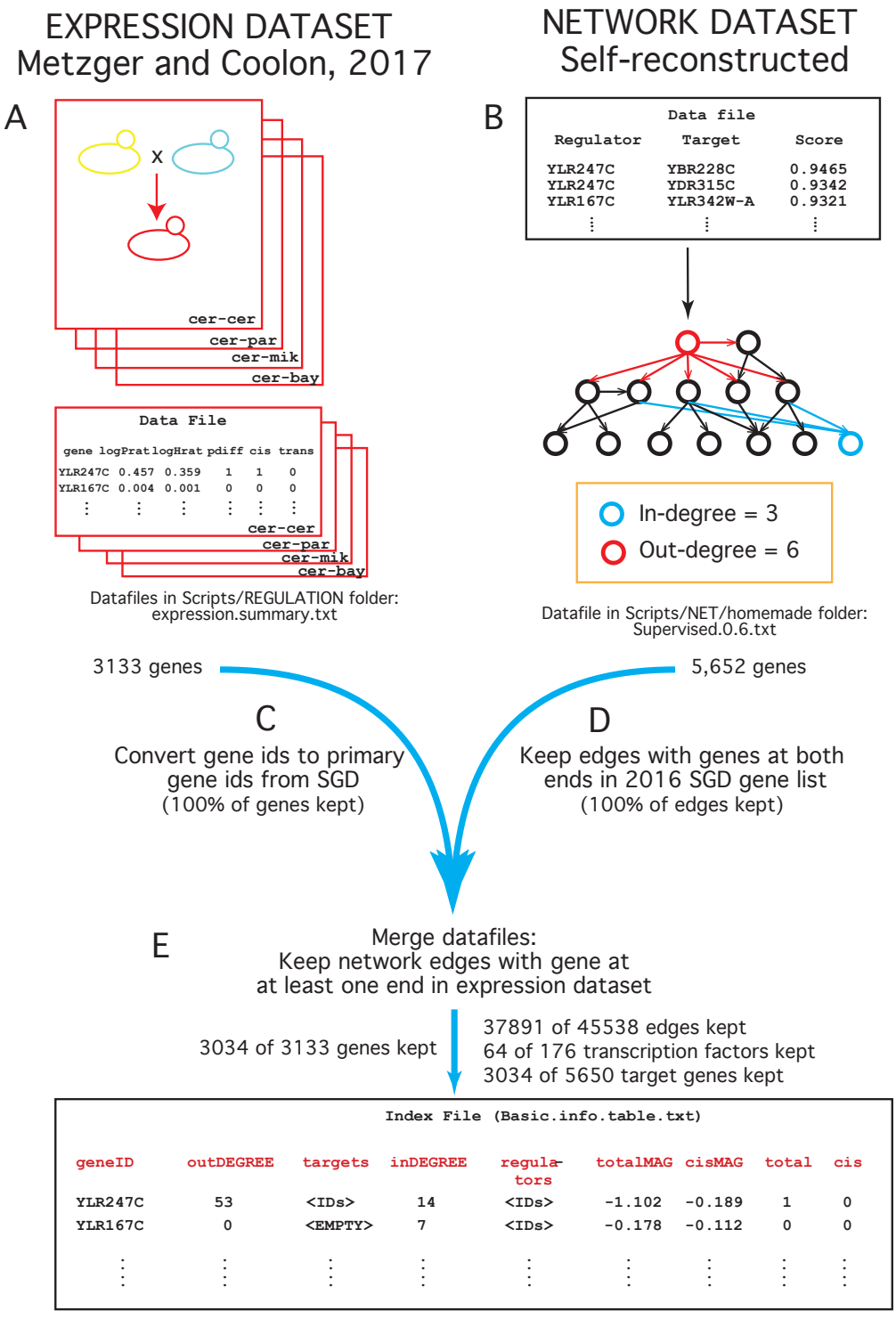

Figure 3.10. Integrating network structure and expression divergence. (**A**) Differences in gene expression and *cis*-regulation between strains and species of *Saccharomyces* were derived from RNAseq data collected from Schaefke et al. 2013 and Schraiber et al. 2013 and analyzed in Metzger et al. 2017. (**B**) The data file describing

the network used in this work was named "Supervised.0.6.txt". For each transcription factor (**Regulator**) - target gene (**Target**) pair, the confidence score (**Score**) describes the probability of the edge. All edges with a probability >0.6 were retained in the network used for our work, which is the same cutoff used by Marbach et al. (2012) for their analyses. A sample network is shown along with the in-degree (# of regulators a target has) of the blue node and out-degree (# of targets a regulator has) of the red node. (**C**) To prepare to merge the gene expression and network datafiles, we converted the gene IDs in the expression data to the primary IDs in the annotation file provided by Saccharomyces Genome Database (SGD). No genes were eliminated at this step. (**D**) Gene names from the network datafile were compared to the annotation file from SGD. All genes in the network datafile have corresponding primary ids. (**E**) The expression and network datafiles were then merged. 3,034 of 3,133 (96.8%) genes in the expression datafile were kept because they appeared at least once in the network. Edges in the network file were kept if the regulator and/or target gene was present in the expression dataset. This filtering retained 37,891 of 45,538 (83.2%) edges, 64 of 176 (36.3%) regulators, and 3,034 of 5,650 (53.6%) targets.

Chapter IV

Existing genetic variants influence properties of mutational effects on gene expression

Abstract

Mutations are the raw input for phenotypic evolution.. However, the extensive epistatic interactions among mutations hinder the efforts to directly connect available genetic variations to the future directions of evolution. Thus, it is important to figure out how the existing genetic variants affect distribution of phenotypic effects of random mutations. In this study, we used the P_{TDH3}-YFP reporter system to specifically study whether the properties of effects on both the mean level of expression and expression noise of random mutations depend on pre-existing genetic variants that perturb different components in the transcriptional program regulating *TDH3* expression. We found that the mean level of expression of the commonly used *Saccharomyces cerevisiae* lab strain BY was more robust to new mutations than other genotypes carrying genetic variants. Also, our results showed that the relationships between the mean level of expression and the expression noise are different in different genetic backgrounds. Finally, we showed that the sensitivity of the mean level of expression to new mutations was positively associated with the expression noise for strains carrying different genetic variants in the promoter. Our results illustrate how various properties of the distribution of effects of random mutations on gene expression changes in response to pre-existing genetic variants in the

genome. Those properties provide foundations for examining predictability of the evolution of gene expression.

Introduction

Genetic variation is the primary source for phenotypic variation. Understanding the phenotypic effects of random mutations is thus important for predicting evolutionary consequences in natural populations. However, until very recently, little empirical evidence has been collected on the phenotypic effects of random mutations. Early studies on the contribution of the phenotypic effects of random mutations to phenotypic evolution focused on using theoretical models to make predictions, one example of which is the Fisher's geometric model (Fisher 1931). One of the important assumptions in the Fisher's geometric model is that the phenotypic effects of new mutations are independent of the underlying genetic background (Martin and Lenormand 2009). Under this assumption, existing mutations in the genome do not affect the evolutionary fate of future mutations. However, it is suggested by multiple studies that the phenotypic effects of new mutations vary in different genetic backgrounds (McKenzie et al. 1982; Remold and Lenski 2004; Milloz et al. 2008; Dworkin et al. 2009; Wang et al. 2013) This phenomenon is called epistasis (Hansen 2013). Theoretical studies have suggested that the widespread epistasis among genes or genetic variations could affect the trajectory of phenotypic evolution (Phillips 2008; Hansen 2013). Previous studies using various systems have shown that the mutations that arise earlier in the genome could completely alter the chance of appearance of new mutations in adaptive evolution (Bridgham et al. 2006; Weinreich et al. 2006; Lang and Desai 2014). Taken together, these theoretical

predictions and empirical studies suggest that it is very important to know what phenotypic effects of new mutations look like in different genetic backgrounds, or more specifically, how do distributions of mutational effects respond to mutations already exist in the genome.

The effects on gene expression of random mutations provide insights on raw inputs for phenotypic evolution, since more and more evidence has suggested that the variation in gene expression is a major contributor to the phenotypic evolution (Stern and Orgogozo 2008; Carroll 2008). One of the interesting discoveries in the recent decades is that not only the mean level of expression within a population, but also the variation in expression level among different individuals, or expression noise, is important for the evolution of gene expression. Gene expression noise refers to the observation that the expression levels among genetically identical individuals in exactly the same environment could be different (Blake et al. 2003). This variation is caused by the molecular fluctuations during each step of gene expression, from transcription, translation to protein degradation (Kaern et al. 2005; Brown et al. 2013). Although the expression noise has been suggested to be important for the phenotypic evolution, it is not clear what evolutionary consequences the expression noise could cause. Theoretical or empirical studies have shown that the expression noise could be either disfavored by selection (under purifying selection) (Fraser et al. 2004; Lehner 2008; Silander et al. 2012; Wolf et al. 2015) or preferred by selection (under positive selection) (Vardi et al. 2013; Zhang et al. 2009). Also, it was found that mutational effects on the expression noise and the mean level of expression were negatively correlated for mutations in *cis* regulatory elements (Metzger et al. 2015). This negative correlation might prevent direct

predictions on the evolutionary trajectory of gene expression, because if selections on the mean level of expression and the expression noise have the same direction, then the optimal level of the expression condition might not be achieved for both metrics, or at least not straightforward to predict. Thus, it is important to know if the relationship between the mean level of expression and the expression noise is the same for random mutations in different genetic backgrounds.

Multiple recent studies have collected data on the effects of either mutations in *cis* elements only or mutations genome wide on gene expression (Metzger et al. 2016; Metzger et al. 2015; Hornung et al. 2012). Due to the important role of the pioneer genetic variants in shaping the future evolutionary paths, it is also necessary to collect data in order to examine whether the existing genetic variants disrupting the regulation of gene expression could also change the properties of mutational effects on both the mean level of expression and the expression noise. In this study, we first examined whether the genetic background influences the mutational effects on both the mean level of expression and the expression noise by mutagenizing two different strains of *Saccharomyces cerevisiae* (YPS1000 and BY, > 53,000 SNPs, 0.44% sequence divergence), each carrying a P_{TDH3}-*YFP* reporter construct in the same genomic location. Then we created 8 strains each carrying a different genetic variant that changes the expression level of the reporter (4 in the promoter, 4 in genes other than *TDH3*) and collected mutational effects on both the mean level of expression and the expression noise in all 8 genetic backgrounds plus the lab strain BY through combination of EMS mutagenesis and flow cytometer phenotyping. We observed that compared to genotypes carrying a single genetic variant, the mean level of the reporter construct in the BY strain

was less sensitive to random mutations, while this is not true for the expression noise. We also found that the relationships between the mutational effects on the mean level of expression and the expression noise were different for different starting genotypes, suggesting that how new mutations simultaneously change the mean level of expression and the expression noise is dependent on pioneer genetic variants already exist in the genome. Finally, we found that for strains carrying genetic variants in the promoter, increasing expression noise was associated with increasing sensitivity of the mean level of expression to random mutations, supporting the hypothesis that the robustness towards molecular fluctuations during gene expression might be related to the robustness to random mutations for gene expression (Lehner 2008; Kaneko 2011). Taken together, our results demonstrate that various aspects of the mutational effects on gene expression are dependent on the pre-existing genetic variants in the genome. And the information we obtained on what the dependences look like could be used to understand the evolutionary fate of individual mutations during the evolution of gene expression.

Results

Genetic background influences effects of random mutations on P_{TDH3}-YFP reporter gene expression

To examine whether and how phenotypic effects of random mutations on gene expression level depend on underlying genetic background, we inserted P_{TDH3}-*YFP* reporter construct in the *HO* locus in two different strains of *Saccharomyces cerevisiae*, BY and YPS1000 respectively. These strains differ by 0.44% in genomic sequence (> 53,000

SNPs). We used a low-dose EMS mutagenesis (Gruber et al. 2012) to collect random mutants (~280 per genetic background) and quantify their reporter expression level through measuring YFP fluorescence signal by using flow cytometer. Effects of mutations on both mean expression level and expression noise were recorded, with expression noise calculated as the coefficient of variation (standard deviation over mean), which is a commonly used measure of expression noise (Kaern et al. 2005). For each genetic background, we also generated a SHAM control population (~140 per genetic background), which experienced all experimental manipulations like mutagenized population except for EMS treatment. All cells within the SHAM population have the same genotype. Variations of expression metrics within SHAM population reflected internal fluctuation of the expression level of the reporter construct. To take this into account, we used z-scores to re-scale the phenotypic effects of random mutations. Z-score provides useful information on how extreme the effect size was for a mutant when compared to the null distribution obtained from SHAM populations. In each genetic background, the z-score for the mean level of expression and/or expression noise of a mutant was calculated as the deviation from average level in the mean level of expression and/or expression noise in SHAM population divided by standard deviation of the SHAM population. All the analysis in this section is presented in both original scale and z-score scale.

The mean level of expression and expression noise level were similar between the two natural strains before mutagenesis (Figure 4.13), suggesting that starting levels of expression metrics did not affect our comparisons. We first compared the number of mutants with either increased or decreased mean expression level between the two strains

(Figure 4.1A and 4.1B). We found that there were similar amounts of random mutations that either increased or decreased expression in BY strain ($N_{increase}$ = 591, $N_{decrease}$ = 622, P-value = 0.389 from Binomial Exact Test). However, the number of mutations that increased mean expression was statistically significantly higher than those that decreased mean expression in the YPS1000 strain ($N_{increase}$ = 172, $N_{decrease}$ = 129, P-value = 0.011 from Binomial Exact Test). Then we compared the range of mutational effects for mean expression level between the two strains. We used Median Absolute Deviation (MAD) instead of standard deviation to measure the range of distributions because MAD was more robust to extreme values. We found that mutations in the YPS1000 strain had a broader range of effects on mean expression level than in BY background (Figure 4.1A, $MAD_{YPS1000}$ = 0.078, MAD_{BY} = 0.047), and analysis using z-scores provided the same conclusion (Figure 4.1B, $MAD_{YPS1000}$ = 0.69, MAD_{BY} = 1.36).

For expression noise, the number of mutations with increasing effect was statistically significant lower than those with decreasing effect (Figure 4.1C, $N_{increase}$ = 555, $N_{decrease}$ = 658, P-value = 0.0033 from Binomial Exact Test), while these two numbers were not significantly different from each other in YPS1000 background (Figure 4.1D, $N_{increase}$ = 166, $N_{decrease}$ = 137, P-value = 0.11 from Binomial Exact Test). Mutations in the YPS strain had a broader range of effects on expression noise than in the BY strain ($MAD_{YPS1000}$ = 0.099, MAD_{BY} = 0.049) in the original scale. However, differences on the range of effect sizes of random mutations between YPS1000 and BY strains decreased ($MAD_{YPS1000}$ = 1.08, MAD_{BY} = 1.06) when used z-score scale. This fact suggested that the larger range of effect sizes observed for YPS1000 strain might be due to the larger variations of expression noise for the starting population.

Taken together, our results suggested that the underlying genetic background affected both direction and range of effect sizes of random mutations. However, because those two strains have > 53,000 differences (0.44% sequence divergence) within their genomic sequences, it is difficult to dissect roles of each genetic variant in modifying mutational effects on the mean level of expression or expression noise. We then moved forward to examine whether mutational effects on expression could change if there was one single genetic variant in the genome that affected expression level.

Mutation rate and measurement of fluorescence were reproducible over time

To examine whether existing genetic variants in the genome could affect the distribution of effects of new mutations, we generated eight strains each carrying a single nucleotide change that was known to affect P_{TDH3}-YFP expression. Among all eight starting strains, four harbored nucleotide changes that were located in the *TDH3* promoter and acted in *cis*. Effects of these *cis* mutations were quantified previously (Metzger et al. 2015). Among these four *cis*-regulatory nucleotide changes, two of them were in known transcription factor binding sites (m76, G -> C in *GCR1* binding site; m66, A -> T in *RAP1* binding site), and the other two were in the TATA box of the *TDH3* gene (TATA1 and TATA2). The other four mutations were located in four different genes in the yeast genome and affected *TDH3* expression in *trans*. Two *trans*-regulatory mutations were in potential regulators of the *TDH3* gene (a non-synonymous mutation in *RAP1*; a non-sense mutation in *TYE7*). *RAP1* has been reported to directly regulate *TDH3* expression (Yagi et al. 1994), and there is indirect evidence in other fungal species suggesting that *TYE7* might also regulate *TDH3* expression (Askew et al. 2009). The other two *trans*-

regulatory changes were located in genes without previously reported regulatory relationship with the *TDH3* gene (a non-synonymous mutation in *ADE6*; a non-synonymous mutation in *NAM7*). All *trans* mutations were identified from a mapping study (Duveau et al. 2014), with the goal to look for *trans* mutations that affected *TDH3* expression. These eight mutations covered a broad range of effect sizes on both the mean level of expression and expression noise (20%-160% on expression mean relative to BY; 50%-300% on expression noise relative to BY) (Figure 4.13). By comparing different aspects of mutational properties responding to the eight starting genetic variants, we could have a better understanding of how differing genetic backgrounds could affect properties of new mutations.

To collect information on mutational effects on reporter gene expression, we used a low dose of EMS to introduce ~40 (31-48, 95% confidence interval) random mutations per cell for each starting genotype. We then used FACS sorting to collect single mutant cell as well as SHAM control cells, and used flow cytometer to quantify reporter construct expression level by measuring fluorescence signals. We collected ~260 mutants and ~130 SHAM control cells for each genotype, and expression level of these cells were used for all downstream analysis.

Due to the scale of the data collection process, we could not mutagenize and quantify all eight starting genotypes on the same day. We used two approaches to check whether results were comparable across experiments. First, for each mutagenesis experiment, we used canavanine assay (Lang and Murray 2008; Gruber et al. 2012) to estimate number of mutations introduced per cell for each genotype (Table 4.1). We found that this number was not significantly different among each other for all starting

genotypes. Thus, differences in distributions of mutational effects were unlikely to be due to amount of mutations introduced. Second, in each mutagenesis experiment, besides mutant genotypes, we also included a BY lab strain and generated its SHAM control population. In total, we had 5 SHAM populations, all with the same genetic background. We compared the distributions of both mean expression level and expression noise among these 5 control populations (Figure 4.2A-B). We used Wilcoxon rank sum test to check if medians of those distributions were significant from each other on both the mean level of expression and expression noise, and also used Kolmogorov-Smirnov two-sample test to check whether each pair of the 5 distributions were from same underlying distribution (Table 4.2 and Table 4.3). We found that measurements of the mean level of expression across experiments for populations from the same genotype were statistically reproducible (Table 4.2). The same was true for expression noise (Table 4.3).

Finally, we did two EMS mutagenesis experiments on M76 genotype on two different days, in order to check whether measurements of distribution on effect sizes of random mutations were robust across different experiments (Figure 4.2C-D). Again, we used Wilcoxon rank sum test and Kolmogorov-Smirnov test to examine similarity on the two distributions of mutational effects from two independent experiments. Both tests suggested that there were no significant differences between measurements from two independent mutagenesis experiments (The mean level of expression, Wilcoxon rank sum test p-value = 0.0031, KS test p-value = 0.0054; Expression Noise, Wilcoxon rank sum test p-value = 0.0097, KS test p-value = 0.014).

Overall, our checking metrics suggested that both number of mutations introduced per cell and expression quantification were reproducible across multiple experiments in different days.

The mean level of expression of wild type lab strain is less sensitive to random mutations than most mutant genotypes

Next, we compared magnitude of mutational effects on the mean level of expression between BY lab strain and all the other mutant genotypes (Figure 4.3). Mutational effects on both the mean level of expression and expression noise of BY lab strain were collected in a previous study (Metzger et al. 2016). We found that magnitude of mutational effects in genetic backgrounds each containing an existing genetic variant were, on average, statistically higher than BY strain (Differences in median mutational effects between BY and other genotypes, Table 4.4; Wilcoxon rank sum test, Table 4.5). To figure out whether this conclusion was true for mutations increasing or decreasing the mean level of expression, we compared mutational effects between BY lab strain and all other mutant genotypes by comparing magnitude of mutational effects for those two types of mutations separately (Figure 4.3B). We found that for both group of mutations, magnitude of mutational effects on the mean level of expression were, on average, statistically significant lower for BY lab strain compared to all the other genotypes (Wilcoxon rank sum test, Table 4.5).

In each mutagenesis experiment, the does of EMS was chosen so that despite multiple mutations (~40, calculated by previously published methods in Gruber et al 2012, see methods) were introduced in each mutant cell, either none or one out of ~40

mutations could have effect size comparable to the effect size quantified from the containing mutant. This potential drawback suggested that directly comparing effect sizes of mutant cells for each starting genetic background might not reflect their true sensitivity to random mutations. To account for this fact, we estimated number of bases (mutational target size) that when mutated, could produce mutational effect on the mean level of expression equal to or greater than a specific cutoff by using methods developed in Metzger et al. (2016). In other words, this metric reflected how many nucleotides in the genome could achieve a specific magnitude of mutational effect when mutated. Comparing mutational target size for different cutoffs across different genotypes could thus give us a measure of how sensitive the starting genotype is to random mutations in the genome.

We thus estimated the mutational target size at different cutoffs for all starting genotypes (Figure 4.4). We found that for majority of the cutoffs, BY lab strain had a smaller mutational target size than other starting genotypes. To better illustrate this, we chose ~300 effect size cutoffs and calculated differences on mutational target size between BY lab strain and each of the mutant genotypes (Figure 4.5). Values larger than zero suggested that the mutant genotype in comparison had a larger mutational target size than BY lab strain for the corresponding effect size cutoff. We found that except for mutations that caused large decreases in TDH3 expression in *RAP1* and *NAM7* mutant genotypes, the BY lab strain had smaller mutational target size than all other mutant genotypes across different effect size cutoffs.

One difference between data for BY lab strain and all other mutant genotypes was the sample size. More than 4 times more random mutants were collected for BY lab

strain than other mutant genotypes. To check how the imbalance of the sample size might affect our interpretations, we generated 200 random samples from BY lab strain data, each containing 290 mutants information randomly selected from 1213 mutants collected for BY strain in a previous study (Metzger et al. 2016). We then repeated estimation of differences in mutational target size on all 200 random samples (Figure 4.6). We found that estimation of mutational target size was less robust for large effect size cutoff, as suggested by the larger 95% confidence interval in tails of each plot (Figure 4.6). Thus, we did not find strong evidence to support the observation that BY lab strain had larger mutational target size in those regions for *RAP1* and *NAM7*. However, majority of the 95% CIs were above zero, suggesting that in majority of the effect size cutoffs, BY lab strain had a significant lower mutational target size than all other mutant genotypes.

We also noticed that distributions of the mean level of expression from SHAM control populations for each starting genotype did not look the same (Figure 4.14A and 4.2B). Similar to what we did for different natural strains, we used z-score to account for the observed different variations within SHAM populations for different starting genotypes (Figure 4.15-4.6). Overall, whey use z-scores, magnitudes of mutational effects from all mutant genotypes were on average, statistically significant higher than BY lab strain (Figure 4.15, Table 4.6 and 4.7). However, when we did similar comparisons separately for mutations increasing or decreasing expression mean, two mutant genotypes showed lower sensitivity than BY lab strain (Table 4.6, 4.7). Mutations increasing expression mean in TATA1 *cis* mutation background and mutations decreasing expression mean in *NAM7 trans* mutation background both showed smaller

magnitude and mutational target size compared to BY (Figure 4.15, 4.5, 4.6). Other than those two cases, BY strain was again found to be statistically less sensitive to random mutations.

Taken together, our results suggest that the mean level of expression of the commonly used lab strain was less sensitive to random mutations than other genotypes with existing mutations disrupting expression level.

Expression noise of wild type lab strain did not show less sensitivity to random mutations than most mutant genotypes

Expression noise affects distribution of gene expression level in natural populations, which is predicted to be important for evolutionary fate of gene expression (Fraser et al. 2004; Lehner 2008; Silander et al. 2012; Wolf et al. 2015). It is thus interesting to ask whether existing genetic variants can also influence mutational effects on expression noise similar to the mean level of expression. We examined whether data for expression noise was consistent with what we found for the mean level of expression (Figure 4.7-4.10). Unlike the mean level of expression, mutational effects on expression noise among 6 out of 8 variant genotypes under investigation did not show significant differences compared to the BY lab strain (Figure 4.7; differences in median mutational effects between BY and each of the other genotypes, results in Table 4.8; Wilcoxon rank sum test on significance for difference recorded in Table 4.8, p-values in Table 4.9). The two starting genotypes carrying the M76 *cis* mutation and the *ADE6 trans* mutation had significantly larger mutational effects on expression noise compared to BY lab strain,

which happened to be the two genetic variants that had the largest effect on expression noise (250% increase for M76; -30% decrease for *ADE6* mutation, Figure 4.13).

We also estimated differences in mutational target size for expression noise on different effect size cutoffs between BY and other genotypes carrying genetic variants (Figure 4.9-4.10). Overall, based on 95% confidence interval estimated from 290 random samples of BY strain (Figure 4.10), we did not find strong evidence to suggest that the BY lab strain had statistically significant different mutational target size compared to other mutant genotypes. Thus, our results suggested that mutational effects on expression noise were not significantly different between the lab strain and mutants carrying genetic variants, except for cases in which the genetic variants had a starting large impact on expression noise..

Disruption of regulatory interactions produced different relationships between the mean level of expression and expression noise

Above we showed that random mutations could affect both expression mean and expression noise. One of the interesting questions about mutational effects on the mean level of expression and expression noise is that whether those two aspects of gene expression have a correlated mutational effect. Previous studies showed that individual *cis* mutations (Hornung et al. 2012; Sharon et al. 2014; Metzger et al. 2015) or *trans* mutations (Metzger et al. 2016) could alter expression mean and noise independently. However, it is not clear whether existing genetic variants in the genome, particularly those that alter direct regulatory interactions, could lead to different relationships between

the mean level of expression and expression noise for random mutations. Here we used mutational data from our 8 mutant genotypes plus BY lab strain to probe this question.

We used Principal Component Analysis (PCA) to decompose primary and secondary axes of variations for each genotype, in order to determine whether there existed co-variation between gene expression mean and noise (Figure 4.11). We also used bootstrap approach to determine whether primary axis of variations in the data was significantly deviate from vertical direction, which suggested no correlation between expression mean and noise (Table 4.10). Interestingly, we found that mutational effects on expression mean and expression noise were significantly different from 90° for four mutant genotypes: mutant carrying M76 *cis* mutation and *ADE6 trans* mutation showed statistically significant negative correlation ($Angle_{M76}$ = 100°, 95% CI [96, 104]; $Angle_{ADE6}$ = 112°, 95% CI [101, 122]), while mutant carrying two TATA box *cis* mutations showed statistically significant positive correlation ($Angle_{TATA1}$ = 83.5°, 95% CI [79, 87]; $Angle_{ADE6}$ = 77.5°, 95% CI [67, 88]). All other genotypes showed no significant relationships between expression mean and expression noise (Table 4.10).

We used z-score data to take into account the variation in SHAM populations for each genotype. We found that all four mutant genotypes showing significant correlation of mutational effects between the mean level of expression and expression noise in original scale also showed same trend using z-scores scale (Figure 4.19, Table 4.11). Thus, our conclusions were robust to specific scales of data used.

Sensitivity to random mutations for the mean level of expression is positively correlated with expression noise for cis-mutant genotypes

Previous studies have suggested that expression noise was proportional to robustness of gene expression level to random mutations (Kaneko 2011; Lehner 2008; Kaneko 2007; Landry et al. 2007). Here, we directly tested this hypothesis by correlating expression noise against variance of mutational effects (Figure 4.12). Specifically, we first estimated average expression noise for each genotype using SHAM control populations. Then we scaled all expression noise values against BY strain expression noise, so that we could get relative size of expression noise for different genotypes. We calculated variance of mutational effects on the mean level of expression for each genotype and correlated those variances against average expression noise (Figure 4.12).

We found that expression noise was positively associated with variance of mutational effects on the mean level of expression for 4 *cis* mutant genotypes and BY strain (P-value = 0.03 from Linear Regression), and was negatively associated for all 4 *trans* mutations. However, since three of the *trans* mutants had very similar variance of mutational effects, it was not clear whether the negative trend was an artifact due to our particular choice of mutant genotypes. Based on these observations, we concluded that for mutant genotypes that differed in promoter sequences, increasing expression noise was associated with increasing sensitivity to random mutations.

Discussion

Although it has long been recognized that the fitness or the phenotypic effects of new mutations depend on the underlying genetic background (McKenzie et al. 1982; Threadgill et al. 1995; Remold and Lenski 2004; Milloz et al. 2008; Dworkin et al. 2009), whether and how a pre-existing genetic variant could affect mutational effects on gene

expression is not clear. Importance of understanding the phenotypic effects of new mutations upon existing genetic variants has been illustrated in multiple studies, all showing that the starting nucleotide changes often dictate the future directions of the evolution in molecular level (Weinreich et al. 2006; Lang and Desai 2014). In this study, we analyzed the mutational effects on both the mean level of gene expression and the expression noise from *Saccharomyces cerevisiae* strains each carrying a single genetic variant. We found that compared to strains carrying genetic variants that disrupt the expression level, the mean level of expression of the BY lab strain was less sensitive to new mutations, but not so for the expression noise. We also found that the existing genetic variants in the genome affect relationship between the mutational effects on the mean level of expression and the expression noise. Finally, we showed that for strains carrying genetic variants in the promoter, the robustness to molecular fluctuations during the process of gene expression was related to the sensitivity to random mutations for the mean level of expression. Below, we discuss connections between our results and current understanding of the mutational effects on gene expression, and implications from our results.

Commonly used wild type strain is more robust to mutational effects on average gene expression level.

The concept of genetic canalization was developed by Waddington in 1942 (Waddington 1942), which referred to the observation that phenotypes of natural populations were robust to the genetic or environmental perturbations. Many experimental systems corroborate the canalization hypothesis since its birth (Gibson and Hogness 1996;

Rutherford and Lindquist 1998; Hall et al. 2007; Braendle and Félix 2008; reviewed in Masel and Siegal 2009; Dworkin 2005). One of the evolutionary consequences of canalization is that variations could accumulate within natural populations without causing strong phenotypic effects, and those mutations are termed cryptic variations (Gibson and Dworkin 2004; Dworkin 2005; Duveau and Félix 2012). One of the predictions from the concept of canalization is that a genetic variant that alters the canalized phenotype might unravel the effects of the cryptic variations. Our data for mutational effects on the mean level of expression is consistent with the prediction from canalization. We found that a single genetic variant that altered the expression level of the *TDH3* gene is enough to cause increased sensitivity to mutational effects on the mean level of expression.

Since canalized phenotypes are robust to genetic perturbations, the adaptation to new environments is restricted due to the lack of available phenotypic variations. Thus, in order to be able to explore a larger mutational space during adaption, pioneer mutations that break the phenotypic robustness might first be fixed within the population so that sensitivity of the phenotype to mutational effects can increase. Multiple studies suggest that existing mutations in the genetic background could affect future path of phenotypic evolution (Weinreich et al. 2006; Bridgham et al. 2006; Bridgham et al. 2006). Our data showed that different starting genetic variants create mutational effects on mean level of expression with different dynamic ranges and symmetry. This fact highlights the importance of understanding how new mutations would respond to the existing genetic variations.

For the expression noise, we did not find statistical significant support that BY strain is less sensitive to new mutations compared to other genotypes (Figure 4.7, Table 4.8-4.9). In fact, there were only two mutant genotypes, M76 and *ADE6* that had stronger mutational effects on expression noise compared to the lab strain BY. Interestingly, those two genetic variants had the strongest impact on expression noise level (Supplementary Figure 1). This fact might suggest that regulatory machinery for expression noise is robust to genetic perturbations, unless those perturbations drastically change the expression noise in the first place.

Our conclusion that the wild type lab strain had a restricted mutational distribution (compared to mutant genotypes) on the mean level of expression while not so on the expression noise is consistent with a previous study on selection strength on mean level of expression and expression noise (Metzger et al. 2015). In the study, the authors found that there was evidence for selection acting on expression noise but not mean level of expression, and this was explained by the exceptional large mutational variance of expression noise of *TDH3* gene in BY strain compared to mean level of expression.

Although consistent with well-documented evolutionary hypothesis, we do notice that both the number of genotypes assayed and the number of mutants collected in the current experiment restrict our power to make solid conclusions. For example, we could not exclude the possibility that other important genetic variants not included in this experiment might show complete opposite trends on the sensitivity to mutational effects. Also, since only ~300 mutants were collected for each genotype, it is possible that we did not explore enough range of the underlying mutational distribution, or the mutational

effects we collected were a biased representation of the underlying mutational distribution. It is necessary to incorporate more genetic variants.

Existing genetic variants can alter relationships between mutational effects on mean level of expression and those on expression noise

Mechanistically speaking, the expression noise is caused by the stochastic fluctuations in the process of gene expression, and it has been illustrated that the stochasticity in transcription is the primary source for the expression noise (Jones et al. 2014). There are two sources of fluctuations during transcription: burst size and burst frequency. Burst frequency refers to how frequent a promoter fires, or switches from transcriptionally inactive state to active state (Elowitz et al. 2002; Brown et al. 2013). The variation in burst size refers to the fact that once promoter is in active state, number of mRNA molecules synthesized in each "burst" varies due to the randomness of RNA polymerase binding and scanning. The mean level of expression describes the average behavior of the gene expression level within a population of genetically identical cells, which averages out fluctuations in both burst frequency and burst size. While the expression noise captures the how large are the fluctuations in both burst frequency and burst size. In other words, burst size and burst frequencies together determine both the mean level of expression and the expression noise. Thus, it is possible that the mutational effects on the mean level of expression and the expression noise are associated with each other..

Theoretical models suggest that increasing expression noise might be deleterious (Fraser et al. 2004), which is supported by the findings that the essential genes have lower expression noise (Silander et al. 2012). However, the expression noise is also

predicted to provide survival benefits in constantly changing environments, due to the existence of individuals with varied phenotypes optimal for different environmental conditions (Vardi et al. 2013). This phenomenon has been linked to the evolvability of gene expression, based on which the expression noise might be beneficial when high evolvability is required (Zhang et al. 2009). Since different models provide different conclusions on the role of the expression noise in the evolution of gene expression, it is not clear whether natural selection would prefer reduction in the expression noise or not. Thus, predicting the evolutionary trajectory of gene expression only based on the mean level of expression without considering simultaneous changes on the expression noise might bias the predictions. In fact, how the mean level of expression and the expression noise relate to each other could completely change the evolutionary consequences on the gene expression. The situation becomes even more complicated based on theories that the selection on the mean level of expression and the expression noise are contradictory to each other in some scenarios (Lehner 2008; Wolf et al. 2015). Also, although it has been suggested that mutations can change the expression noise and the mean level of expression independently (Thattai and Van Oudenaarden 2001; Munsky et al. 2012; Hornung et al. 2012; Murphy et al. 2010), it was also found that there was a negative correlation between the mutational effects on the expression noise and the mean level of expression for mutations in *cis* (Metzger et al. 2016). Thus, knowing whether and how the mutational effects on the expression noise and the mean level of expression vary with each other in different genetic backgrounds help us gain new insights on the potential constraints on the evolution of gene expression.

In our analysis, we found that new mutations in different starting genotypes could generate different relationships between the mean level of expression and the expression noise. Specifically, we found that the two strains with genetic variants in TATA box in the *TDH3* promoter both showed positive correlations between the mutational effects on the mean level of expression and the expression noise, whereas the two genetic variants with largest effects on the expression level of *TDH3* gene, one in transcription factor binding sites for GCR1 (M76) and the other in *ADE6* gene, generated negative correlations between the mutational effects on the mean level of expression and the expression noise. It is suggested that the expression noise is controlled by nucleosome positioning in the TATA box (Murphy et al. 2010; Hornung et al. 2012). Thus, genetic variants that disrupt the TATA box might create specific patterns of correlation between mutational effects on the mean level of expression and the expression noise. Despite unexplained molecular mechanism, our data suggest that introducing different genetic variants into the genome could result in different relationships between the mutational effects on the mean level of expression and the expression noise.

One prediction of our result is that if selection prefers changes in the mean level of expression and the expression noise in the same direction, genetic variants that result in positive correlation between the mutational effects on the two metrics might be preferentially fixed in the population, because more fitted mutations could be fixed following the fixation of those genetic variants The opposite would happen if selection prefers changes in the mean level of expression and the expression noise in opposite directions. From this perspective, our findings provide insights on predictability of genetic basis for the evolution of gene expression in the molecular level.

Robustness to internal fluctuation is informative for robustness to random mutations

Theoretical studies suggest that the expression noise is proportional to sensitivity of phenotypes to random mutations (Kaneko 2011). In Lehner (2008), the author used data from a yeast mutation accumulation study (Landry et al. 2007) and found that genes whose expression level were less sensitive to spontaneous mutations have smaller expression noise, and those genes tend to be essential genes. Thus, selection to minimize the expression noise for essential genes might be linked to the canalization of the expression level towards genetic perturbations.

In the current study, we showed that the expression noise of strains carrying a genetic variant in the promoter region of *TDH3* reporter construct was positively correlated with variance of the distributions of mutational effects on the mean level of expression. Our conclusion implies that disrupting the *cis* regulatory elements might have coordinated effects on both the expression noise and the sensitivity to random mutations, which is consistent with Lehner's study. Taken together, both studies suggested that there might exist some mechanisms that link the expression noise to the sensitivity towards random mutations on the mean level of expression. However, we do notice that the number of genetic variants assayed in the current experiments limits our power to draw strong conclusions. With only four data points, we could not exclude the possibility that the observed positive correlation might be due to the bias in selecting the genetic variants.

However, we did not find the same relationship between the expression noise and variation in the mutational effects on the mean level of expression for all four *trans*

genetic variants. However, we think our power to draw solid conclusion for *trans* genetic variants is limited, because three of the four *trans* genetic variants had similar level of the variations of mutational effects. That said, we effectively only have two data points in our analysis for *trans* genetic variants. This fact suggests that more *trans* genetic variants are required in order to get more accurate conclusions on relationship between the robustness to molecular fluctuations during transcription and the robustness to genetic perturbations.

One prediction from our immature conclusion is that the direction of selection on the expression noise might be related to the selection pressure on the sensitivity to random mutations. If the robustness to genetic variations for phenotypes is preferred by selection, than the genetic variants that result in smaller expression noise might be preferentially fixed since they convey smaller sensitivity to random mutations. This is another example that our results are useful in understanding the potential evolutionary fate of individual mutations.

Future remarks

In this study, we analyzed how existing mutations in the genome could affect the properties of mutational effects on both the mean level of expression and the expression noise. As discussed above, those properties are important for a better description of the evolutionary trajectory of the evolution of gene expression in molecular level. We propose that to examine the generality of our conclusions from a system on a single gene and limited number of genetic variants, more experimental systems could be designed to explore more types of genetic variants and how they affect the effects of random

mutations. Also, the information we obtained on how existing genetic variants influenced the mutational effects could be incorporated into mathematical models to better understand the process of the evolution of gene expression.

Materials and Methods

Genetic background of yeast strains

All mutant strains were created in strain YPW1139 (*Matα*). This strain is derived from BY4724, BY4722, BY4730 and BY4742 and contains no auxotrophies. This strain also contains five mutations from natural yeast strains that fix two common defects in laboratory strains: high frequency of petites and low sporulation rate. Finally, this strain contains a copy of the *TDH3* promoter from BY laboratory strain, YFP coding sequence, *CYC1* terminator, and KanMX4 drug resistance cassette inserted at the *HO* locus on chromosome IV. Further details about this strain can be found in other studies published from Wittkopp lab (Metzger et al. 2015; Duveau et al. 2014; Metzger et al. 2016). All mutant strains carrying *cis*-regulatory mutations were previously created using site directed mutagenesis approach, as described in Metzger and Yuan. (2015). Each *cis* mutation was inserted into the *TDH3* promoter of the reporter cassette in *HO* locus. *Trans* mutations in *ADE6, TYE7* and *NAM7* were identified in Duveau et al. (2014). *Trans* mutation in *RAP1* was identified by a systematic PCR mutagenesis approach using random primers. All mutant strains carrying *trans*-regulatory mutations were created in YPW1139 strain, using site directed mutagenesis approach described in Duveau et al. (2014). Each *trans* mutation was inserted into its native genomic location. The natural

S.cerevisiae strains YPS1000 (PJW1057) containing the P_{TDH3}-YFP cassette in *HO* locus was generated in the same way as described above.

Mutagenesis

To generate *trans*-regulatory mutations in each genotype, a previously published EMS mutagenesis protocol was used, as described in Gruber et al. (2012). The specific dose of EMS used in the protocol was chosen so that the proportion of treated cells carrying single mutation that significantly changed expression level of reporter gene was maximized, while proportion of cells carrying no causal mutations was minimized (Gruber et al. 2012). Experiments and calculations to illustrate criteria for choosing EMS dosage could be found in Metzger et al. (2016). Before each mutagenesis experiment, yeast cells of all genotypes used in the experiment were revived from glycerol stock stored in -80 °C freezer by plating cells on YPG agar medium (10g/L yeast extract, 20g/L peptone, 5% vol/vol glycerol and 20g/L agar), in order to minimize number of petites. After ~2 days, cells on plates were transferred into 10ml YPD liquid (20g/L mono-saccharides, 10g/L yeast extract and 20g/L peptone) and cultured for 10-11 cell cycles (~24 hours) under 30 °C and ~200rpm shaking. Before mutagenesis, cells were washed in 1ml 1X PBS and 1ml H_2O for twice, and re-suspended in 1ml of sodium phosphate (0.1M). To mutagenize, cells were treated with 10ul EMS, achieving a final concentration of 1%. After 45 minutes, EMS was quenched with 1ml 5% sodium thiosulfate, and cells were washed twice each using 1ml 5% sodium thiosulfate and 1ml H_2O in Eppendorf 1.7ml centrifuge tubes. After the two wash steps, cells were washed and suspended in 1ml YPD liquid, and 0.125ml from the resulted culture was transferred

to 3.875 ml fresh YPD liquid medium. The SHAM control population of cells for each genotype experienced all wash and dilution steps like EMS treated cells, except for EMS treatment. After 24 hours, 0.125 ml of 4ml cultures from previous step was diluted into 3.875 ml fresh YPD liquid medium, in order to avoid saturation of yeast growth. The purpose of the two dilutions and growth afterwards was to recover cells from EMS treatment stress (for ~10 cell cycles).

To control for potential bias caused by differing number of mutations introduced in each experiment, mutation rate was calculated using a canavanine resistant assay (Gruber et al. 2012). Specifically, for each EMS-treated culture, 0.1 ml of 10^{-1} dilution of cells were plated on agar medium (7.1g bacto-yeast nitrogen base, 20g/L dextrose, 2g/L amino acid mix without arginine and 20g/L agar) with 60mg/ml L-canavanine sulfate and without arginine. In parallel, 0.1 ml of $2x10^{-4}$ dilutions of cells were also plated on same agar medium but without L-canavanine. Colonies formed on first plates contained cells carrying canavanine resistance, and number of colonies on second plates was used to normalize against concentration of the culture, as described in Gruber et al. (2012). Previous study (Lang and Murray 2008) showed that there were 88 EMS-like mutations in yeast genome that can result in canavanine resistance. Based on number of colonies on the two types of medium and number of causal mutations for canavanine resistance, mutation rate was calculated for each mutagenesis experiment, achieving on average ~40 mutations per genome per experiment (see Supplementary Table 4.1).

Sorting mutants using FACS

After 48 hours recovery from EMS treatment, individual EMS treated and SHAM yeast cells were plated in 384-well format on YPD agar medium using FACS (BD FACS Aria II, University of Michigan Flow Cytometer Core). Before FACS sorting, 0.5ml YPD liquid culture (~1x10^7 cells) was diluted in 1.5ml 1X PBS solution. Each sample was then run on the FACS machine at a flow rate of ~15000 cells/s. During sorting, non-yeast events or cell aggregates were removed through setting gating on flow cytometer using FACSDiva software. For each genotype, 300 EMS treated cells and 150 SHAM control cells were collected on YPD agar medium.

After sorting, cells were cultured on YPD agar medium for ~48 hours at 30 °C. On the plate, no growth was observed for ~6%-10% of positions, which might due to lethal mutations or non-sorting event. This resulted in ~270 EMS treated cells and ~130 SHAM control cells survived for all downstream analysis in each experiment. After 48 hours culture, 4 quadrants of 96 colonies were transferred to 4 96-well plates with V&P pin tool, and those plates were filled with 0.5ml YPD liquid medium in each well beforehand. In parallel, YPW1139 was revived from glycerol stock in -80 °C freezer and plated on 24 fixed positions in the 96-well plates described above. Those cells were used to correct potential plate effect during culture for downstream fluorescence analysis. Note that those "calibration" cells were not through neither mutagenesis treatment nor single cell bottleneck. In addition, in each plate, 2 copies of strain YPW978 carrying no YFP reporter in the genome were placed in two random positions, in order to remove auto-fluorescence signal in downstream analysis. All 96-well plates were cultured under 30 °C and shaking (~500rpm) for 24 hours. In the next day, 100ul YPD culture from each well in 96-well plates were mixed with 23ul 80% glycerol to make glycerol stock in

96-well format in -80 °C freezer. At the same time, cultures from all wells of each 96-well plate were transferred onto an YPG agar medium using pin tool, in order to remove petites. Overall, ~0.5%-2% petites were observed for each experiment. Cells were cultured on YPG agar medium for another 48 hours and then transferred to 4 replicate 96-well plates containing 0.5ml YPD liquid medium in each well before fluorescent measure on BD Accuri C6 machine. The number of cells for each genotype in each experiment is summarized in Supplementary Table 2.

Phenotyping using BD Accuri C6

After culture in YPD in 96-well plates for 22-24 hours, 13-15ul liquid cultures were then mixed with 0.5ml 1X PBS in each well in another 96-well plate. Fluorescence was recorded by using HyperCyt autosampler (Intellicyt Corp) connected to a BD Accuri C6 machine (488nm laser for fluorescence excitation and 533/30nm optical filter for signal acquisition). For each sample, ~$2x10^4$ fluorescence events were recorded.

Flow data analysis

Flow cytometry data were analyzed by using previously published methods with several modifications (Metzger et al. 2016), with brief steps summarized as follows. First, clustering functions in R package *flowClust* (Lo et al. 2009) and *flowCore* (Ellis et al. 2009) were used to filter out all events that did not record single-cell events based on height and area of forward scatter signal. Then, intensity of fluorescence signal was scaled by cell size through multiple steps, the purpose of which is to ensure correction of fluorescence signal was robust to different relationships between fluorescence signal and

cell size observed in different genotypes and environment. Next, the re-scaled fluorescence signal is transformed with a function log (new value) = 10.469 x log (scaled fluorescence) – 9.586, so that the transformed value is linearly related to the YFP mRNA level. This formula is obtained from (Duveau, unpublished data), in which Duveau used both pyro-sequencing to quantify YFP mRNA abundance and flow cytometer to quantify YFP protein activity, and interpolated formula from (Wang and Gaigalas 2011) by the two types of data obtained to quantify the relationship between fluorescence and mRNA level. Samples with less than 2000 events after all the above filtering steps were removed due to inadequate amount of data for downstream analysis. Median and standard deviation of transformed fluorescence in each well (sample) were calculated and recorded. Then, medians or standard deviations of transformed fluorescence of the 24 wells containing YPW1139 cells not experiencing EMS treatment or wash steps were used to correct for technical factors that change fluorescence without genetic basis. These factors include positions on the plate, positions in the tower used for shaking plates, access to oxygen during culture and measurement fluctuation by the Accuri machine. Mixed linear model (Expression ~ 1 + RUN + (ROW | RUN), where RUN is an integer number representing # of Accuri run, and ROW is an integer number representing row # on 96 well plates) was used for corrections on both median and standard deviation of transformed fluorescence and corrected values were retained for downstream analysis. Expression mean and standard deviation for each mutant or SHAM genotype were then calculated as mean of the four replicates. Expression noise was calculated as coefficient of variation (average standard deviation across replicates over average mean).

Estimating Mutational Target Size

Methods in Metzger et al. (2016) were used to calculate mutational target size. In brief, for each genotype, the proportion of *trans*-mutants among all collected *trans*-mutants with effect size equal or greater than a specific cutoff was used to calculate number of random mutations with effect size equal or greater than the same cutoff. For example, if 30 out of 300 mutants had effect sizes equal or greater than +2%, then, assuming total number of nucleotides in yeast genome was ~12000000, the expected number of mutations with effect size >= +2% was calculated as 12000000 x 30 / 300. Since each *trans*-mutant contained on average 40 mutations, as estimated from canavanine assay, this fact should be corrected in calculating potential target size. Two assumptions were made for the calculation. First, within each mutant, no mutation had effect size larger than effect size of the mutant measured empirically. In another word, no large effect, second-site, compensatory mutations were present in any mutant. Second, the number of mutations with a specific effect size within a mutant followed a Poisson distribution. Under this assumption, proportion of mutants below a specific effect cutoff would be proportional to $e^{-\lambda}$, where λ is the Poisson rate parameter. The λ parameter was estimated using the method described above and bias caused by existence of multiple mutations were corrected using Poisson distribution. As was suggested in Metzger et al. (2016), this bias is large for mutants with small effect and trivial for mutants with large effects.

Statistical Analysis

All statistical analysis presented in main text were performed on R (version 3.2.2, R Core Team 2015). Scripts could be found in Deepblue.

References

Askew C, Sellam A, Epp E, Hogues H, Mullick A, Nantel A, Whiteway M. 2009. Transcriptional Regulation of Carbohydrate Metabolism in the Human Pathogen Candida albicans.Andrianopoulos A, editor. PLoS Pathog 5:e1000612.

Blake WJ, Kaern M, Cantor CR, Collins JJ. 2003. Noise in eukaryotic gene expression. Nature 422:633–637.

Braendle C, Félix M-A. 2008. Plasticity and errors of a robust developmental system in different environments. Dev. Cell 15:714–724.

Bridgham JT, Carroll SM, Thornton JW. 2006. Evolution of Hormone-Receptor Complexity by Molecular Exploitation. Science 312:97–101.

Brown CR, Mao C, Falkovskaia E, Jurica MS, Boeger H. 2013. Linking Stochastic Fluctuations in Chromatin Structure and Gene Expression.Rando OJ, editor. PLOS Biol 11:e1001621.

Carroll SB. 2008. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. Cell 134:25–36.

Duveau F, Félix M-A. 2012. Role of Pleiotropy in the Evolution of a Cryptic Developmental Variation in Caenorhabditis elegans.Noor MAF, editor. PLOS Biol 10:e1001230.

Duveau F, Metzger BPH, Gruber JD, Mack K, Sood N, Brooks TE, Wittkopp PJ. 2014. Mapping small effect mutations in Saccharomyces cerevisiae: impacts of experimental design and mutational properties. G3 (Bethesda) 4:1205–1216.

Dworkin I, Kennerly E, Tack D, Hutchinson J, Brown J, Mahaffey J, Gibson G. 2009. Genomic Consequences of Background Effects on scalloped Mutant Expressivity in the Wing of Drosophila melanogaster. Genetics 181:1065–1076.

Dworkin I. 2005. Canalization, cryptic variation, and developmental buffering: a critical examination and analytical perspective. Variation: A central concept in biology.

Ellis B, Haaland P, Hahne F, Le Meur N. 2009. flowCore: basic structures for flow cytometry data. R package version

Elowitz MB, Levine AJ, Siggia ED, Swain PS. 2002. Stochastic gene expression in a single cell. Science.

Fraser HB, Hirsh AE, Giaever G, Kumm J, Eisen MB. 2004. Noise minimization in eukaryotic gene expression.Ken Wolfe, editor. PLOS Biol 2:e137.

Fisher., 1931. The Genetic Theory of Natural Selection.

Gibson G, Dworkin I. 2004. Uncovering cryptic genetic variation. Nature Reviews Genetics 5:681–690.

Gibson G, Hogness DS. 1996. Effect of polymorphism in the Drosophila regulatory gene Ultrabithorax on homeotic stability. Science 271:200–203.

Gruber JD, Vogel K, Kalay G, Wittkopp PJ. 2012. Contrasting properties of gene-specific regulatory, coding, and copy number mutations in Saccharomyces cerevisiae: frequency, effects, and dominance.Akey JM, editor. PLoS Genet. 8:e1002497.

Hall MC, Dworkin I, Ungerer MC, Purugganan M. 2007. Genetics of microenvironmental canalization in Arabidopsis thaliana. Proceedings of the National Academy of Sciences 104:13717–13722.

Hansen TF. 2013. WHY EPISTASIS IS IMPORTANT FOR SELECTION AND ADAPTATION. Evolution 67:3501–3511.

Hornung G, Bar-Ziv R, Rosin D, Tokuriki N, Tawfik DS, Oren M, Barkai N. 2012. Noise-mean relationship in mutated promoters. Genome Res. 22:2409–2417.

Jones DL, Brewster RC, Phillips R. 2014. Promoter architecture dictates cell-to-cell variability in gene expression. Science.

Kaern M, Elston TC, Blake WJ, Collins JJ. 2005. Stochasticity in gene expression: from theories to phenotypes. Nature Reviews Genetics 6:451–464.

Kaneko K. 2007. Evolution of robustness to noise and mutation in gene expression dynamics.Scalas E, editor. PLoS ONE 2:e434.

Kaneko K. 2011. Proportionality between variances in gene expression induced by noise and mutation: consequence of evolutionary robustness. BMC Evol. Biol. 11:27.

Landry CR, Lemos B, Rifkin SA, Dickinson WJ, Hartl DL. 2007. Genetic Properties Influencing the Evolvability of Gene Expression. Science 317:118–121.

Lang GI, Desai MM. 2014. The spectrum of adaptive mutations in experimental evolution. Genomics 104:412–416.

Lang GI, Murray AW. 2008. Estimating the Per-Base-Pair Mutation Rate in the Yeast

Saccharomyces cerevisiae. Genetics 178:67–82.

Lehner B. 2008. Selection to minimise noise in living systems and its implications for the evolution of gene expression. Molecular Systems Biology 4:170.

Lo K, Hahne F, Brinkman RR, Gottardo R. 2009. flowClust: a Bioconductor package for automated gating of flow cytometry data. BMC Bioinformatics 10:145.

Martin G, Lenormand T. 2009. A GENERAL MULTIVARIATE EXTENSION OF FISHER'S GEOMETRICAL MODEL AND THE DISTRIBUTION OF MUTATION FITNESS EFFECTS ACROSS SPECIES. http://dx.doi.org/10.1554/05-412.1 60:893.

Masel J, Siegal ML. 2009. Robustness: mechanisms and consequences. Trends in Genetics 25:395–403.

McKenzie JA, Whitten MJ, Adena MA. 1982. The effect of genetic background on the fitness of diazinon resistance genotypes of the Australian sheep blowfly, Lucilia cuprina. Heredity 49:1–9.

Metzger BPH, Duveau F, Yuan DC, Tryban S, Yang B, Wittkopp PJ. 2016. Contrasting Frequencies and Effects of cis- and trans-Regulatory Mutations Affecting Gene Expression. Mol. Biol. Evol. 33:1131–1146.

Metzger BPH, Yuan DC, Gruber JD, Duveau F, Wittkopp PJ. 2015. Selection on noise constrains variation in a eukaryotic promoter. Nature 521:344–347.

Milloz J, Duveau F, Nuez I, Félix M-A. 2008. Intraspecific evolution of the intercellular signaling network underlying a robust developmental system. Genes Dev. 22:3064–3075.

Munsky B, Neuert G, Van Oudenaarden A. 2012. Using gene expression noise to understand gene regulation. Science.

Murphy KF, Adams RM, Wang X, Balázsi G, Collins JJ. 2010. Tuning and controlling gene expression noise in synthetic gene networks. Nucl. Acids Res. 38:2712–2726.

Phillips PC. 2008. Epistasis |[mdash]| the essential role of gene interactions in the structure and evolution of genetic systems. Nature Reviews Genetics 9:855–867.

Remold SK, Lenski RE. 2004. Pervasive joint influence of epistasis and plasticity on mutational effects in Escherichia coli. Nature Genetics 36:423–426.

Rutherford SL, Lindquist S. 1998. Hsp90 as a capacitor for morphological evolution. Nature 396:336–342.

Sharon E, van Dijk D, Kalma Y, Keren L, Manor O, Yakhini Z, Segal E. 2014. Probing the effect of promoters on noise in gene expression using thousands of designed

sequences. Genome Res. 24:1698–1706.

Silander OK, Nikolic N, Zaslaver A, Bren A, Kikoin I, Alon U, Ackermann M. 2012. A genome-wide analysis of promoter-mediated phenotypic noise in Escherichia coli.Matic I, editor. PLoS Genet. 8:e1002443.

Stern DL, Orgogozo V. 2008. THE LOCI OF EVOLUTION: HOW PREDICTABLE IS GENETIC EVOLUTION? Evolution 62:2155–2177.

Thattai M, Van Oudenaarden A. 2001. Intrinsic noise in gene regulatory networks.

Threadgill DW, Yee D, Thompson C, Magnuson T. 1995. Epidermal Growth Factor Receptor Deficiency Results in Periimplantation Lethality in Mouse. In: Molecular and Cellular Aspects of Periimplantation Processes. New York, NY: Springer New York. pp. 231–235.

Vardi N, Levy S, Assaf M, Carmi M, Barkai N. 2013. Budding yeast escape commitment to the phosphate starvation program using gene expression noise. Curr. Biol. 23:2051–2057.

Waddington CH. 1942. Canalization of Development and the Inheritance of Acquired Characters. Nature 150:563–565.

Wang L, Gaigalas AK. 2011. Development of Multicolor Flow Cytometry Calibration Standards: Assignment of Equivalent Reference Fluorophores (ERF) Unit. J Res Natl Inst Stand Technol 116:671–683.

Wang Y, Arenas CD, Stoebel DM, Cooper TF. 2013. Genetic background affects epistatic interactions between two beneficial mutations. Biology Letters 9:20120328–20120588.

Weinreich DM, Delaney NF, DePristo MA, Hartl DL. 2006. Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. Science 312:111–114.

Wolf L, Silander OK, van Nimwegen E. 2015. Expression noise facilitates the evolution of gene regulation. Elife 4:987.

Yagi S, Yagi K, FUKUOKA J, SUZUKI M. 1994. The UAS of the yeast GAPDH promoter consists of multiple general functional elements including RAP1 and GRF2 binding sites. The Journal of Veterinary Medical Science 56:235–244.

Zhang Z, Qian W, Zhang J. 2009. Positive selection for elevated gene expression noise in yeast. Molecular Systems Biology 5:299.

Table 4.1. Number of mutants introduced per mutant cell for each genotype estimated from canavanine assay. 95% confidence interval was calculated from 6 replicates.

Genotype	Estimated number	95% CI
M76	**39**	**(31, 47)**
TATA1	**41**	**(36, 46)**
M66	**40**	**(33, 47)**
TATA2	**37**	**(31, 43)**
RAP1	**38**	**(30, 46)**
ADE6	**39**	**(31, 47)**
TYE7	**34**	**(28, 40)**
NAM7	**36**	**(29, 43)**

Table 4.2. Comparisons of the mean level of expression for the 5 SHAM populations by Wilcoxon rank sum test and KS test. Results from pairwise comparisons of statistical significant differences among 5 BY SHAM populations across multiple mutagenesis experiments for the mean level of expression. In each parenthesis, first number is pvalue from Wilcoxon rank sum test and second number is pvalue from Kolmogorov-Smirnov two-sample test. All pvalues were adjusted by Hochberg-Benjamin multiple test correction procedure.

Experiments	BY1	BY2	BY3	BY4	BY5
BY1		(0.42, 0.48)	(0.63, 0.27)	(0.96, 0.71)	(0.20, 0.11)
BY2			(0.29, 0.26)	(0.47, 0.21)	(0.13, 0.15)
BY3				(0.64, 0.52)	(0.14, 0.21)
BY4					(0.09, 0.12)
BY5					

Table 4.3. Comparisons of the expression noise for the 5 SHAM populations by Wilcoxon rank sum test and KS test. Results from pairwise comparisons of statistical significant differences among 5 BY SHAM populations across multiple mutagenesis experiments for expression noise. In each parenthesis, first number is pvalue from Wilcoxon rank sum test and second number is pvalue from Kolmogorov-Smirnov two-sample test. All pvalues were adjusted by Hochberg-Benjamin multiple test correction procedure.

Experiments	BY1	BY2	BY3	BY4	BY5
BY1		(0.96, 0.26)	(0.48, 0.15)	(0.21, 0.12)	(0.09, 0.08)
BY2			(0.75, 0.21)	(0.44, 0.30)	(0.06, 0.12)
BY3				(0.80, 0.58)	(0.11, 0.17)
BY4					(0.04, 0.11)
BY5					

Table 4.4. Differences the average magnitude of mutational effects on the mean level of expression between BY strain and each of the other genetic backgrounds. This difference was calculated separately for all mutants (1st row), mutants increasing mean level expression (2nd row) and mutants decreasing the mean level of expression (3rd row). Results were calculated by subtracting the average magnitude of mutational effects of BY strain from the corresponding value of other genetic backgrounds. If the value is larger than zero, then random mutations have on average higher magnitude of effect in strains carrying genetic variants than BY strain.

CAT	M76	TATA1	M66	TATA2	RAP1	ADE6	TYE7	NAM7
ALL	0.059	0.023	0.015	0.011	0.007	0.026	0.006	0.010
INCREASE	0.057	0.022	0.021	0.012	0.011	0.018	0.007	0.016
DECREASE	0.061	0.026	0.009	0.009	0.004	0.036	0.006	0.008

Table 4.5. Comparisons of magnitude of mutational effects on the mean level of expression between BY strain and all other strains with existing genetic variants. Statistical significances of differences on mutational effect sizes between BY and each of the other genotypes were examined using Wilcoxon rank sum test. All pvalues were from Wilcoxon rank sum test and adjusted by Hochberg-Benjamin multiple test correction procedure.

CAT	M76	TATA1	M66	TATA2	RAP1	ADE6	TYE7	NAM7
ALL	1.5×10^{-67}	1.5×10^{-33}	9.1×10^{-16}	3.1×10^{-11}	1.0×10^{-8}	5.9×10^{-40}	2.8×10^{-5}	1.5×10^{-12}
INCREASE	1.0×10^{-32}	8.0×10^{-14}	3.1×10^{-15}	5.3×10^{-7}	6.6×10^{-10}	4.7×10^{-13}	5.0×10^{-5}	1.2×10^{-9}
DECREASE	7.4×10^{-36}	5.5×10^{-20}	1.1×10^{-3}	1.9×10^{-6}	0.061	2.4×10^{-28}	0.063	2.8×10^{-4}

Table 4.6. Differences in the average magnitude of mutational effects (in z-score scale) on the mean level of expression between BY strain and each of the other genetic backgrounds. The differences were calculated separately for all mutants (1st row), mutants increasing mean level expression (2nd row) and mutants decreasing the mean level of expression (3rd row). Results were calculated by subtracting the average magnitude of mutational effects of BY strain from the corresponding value of other genetic backgrounds. If the value is larger than zero, then random mutations have on average higher magnitude of effect in strains carrying genetic variants than BY strain.

CAT	M76	TATA1	M66	TATA2	RAP1	ADE6	TYE7	NAM7
ALL	0.39	0.24	0.85	0.43	0.38	0.52	0.58	0.08
INCREASE	0.37	-0.08	1.19	0.48	0.54	0.30	0.61	0.22
DECREASE	0.40	0.30	0.56	0.36	0.26	0.78	0.56	-0.02

Table 4.7. Comparisons of magnitude of mutational effects (in z-score scale) on the mean level of expression between BY strain and all other strains with existing genetic variants. Statistical significances of differences on mutational effect sizes between BY and each of the other genotypes were examined using Wilcoxon rank sum test. All pvalues were from Wilcoxon rank sum test and adjusted by Hochberg-Benjamin multiple test correction procedure.

CAT	M76	TATA1	M66	TATA2	RAP1	ADE6	TYE7	NAM7
ALL	2.2×10^{-24}	3.6×10^{-21}	8.4×10^{-23}	5.5×10^{-19}	2.2×10^{-9}	7.9×10^{-31}	4.4×10^{-9}	2.3×10^{-11}
INCREASE	2.2×10^{-10}	0.021	4.5×10^{-8}	1.4×10^{-9}	7.7×10^{-6}	2.5×10^{-5}	5.0×10^{-6}	3.3×10^{-7}
DECREASE	1.1×10^{-6}	4.5×10^{-9}	1.2×10^{-11}	1.8×10^{-12}	6.4×10^{-5}	3.6×10^{-28}	4.3×10^{-8}	0.012

Table 4.8. Differences in the average magnitude of mutational effects on expression noise between BY strain and each of the other genetic backgrounds. The differences were calculated separately for all mutants (1st row), mutants increasing mean level expression (2nd row) and mutants decreasing the mean level of expression (3rd row). Results were calculated by subtracting the average magnitude of mutational effects of BY strain from the corresponding value of other genetic backgrounds. If the value is larger than zero, then random mutations have on average higher magnitude of effect in strains carrying genetic variants than BY strain.

CAT	M76	TATA1	M66	TATA2	RAP1	ADE6	TYE7	NAM7
ALL	0.07	0.01	0.004	0.001	0.006	0.03	$7x10^{-5}$	0.014
INCREASE	0.13	0.013	0.006	0.002	0.004	0.043	0.004	0.013
DECREASE	0.014	0.009	0.003	0.001	0.007	0.019	-0.001	0.012

Table 4.9. Comparisons of magnitude of mutational effects on expression noise between BY strain and all other strains with existing genetic variants. Statistical significances of differences on mutational effect sizes between BY and each of the other genotypes were examined using Wilcoxon rank sum test. All pvalues were from Wilcoxon rank sum test and adjusted by Hochberg-Benjamin multiple test correction procedure.

CAT	M76	TATA1	M66	TATA2	RAP1	ADE6	TYE7	NAM7
ALL	5.7×10^{-43}	9.4×10^{-5}	0.88	0.54	0.06	1.6×10^{-18}	0.53	1.3×10^{-4}
INCREASE	1.7×10^{-47}	7.3×10^{-3}	0.69	0.57	0.23	1.1×10^{-14}	0.78	0.012
DECREASE	3.1×10^{-5}	0.005	0.77	0.89	0.064	2.1×10^{-5}	0.23	0.004

Table 4.10. Angles of the primary axis of variation estimated from principal component analysis on the mean level of expression against expression noise for all genetic backgrounds. Statistical significance on whether the estimated angles were different from zero was estimated using bootstrap approach. P-values were estimated from 10000 bootstrap samples.

Genotype	Angle	P-value	95% CI
WT	**97°**	**0.13**	**(85, 109)**
M76	**100°**	**< 0.00001**	**(96, 104)**
TATA1	**83.5°**	**0.013**	**(79, 87)**
M66	**88.5°**	**0.69**	**(18, 159)**
TATA2	**77.5°**	**0.0001**	**(67, 88)**
RAP1	**90°**	**0.60**	**(81, 99)**
ADE6	**111.5°**	**< 0.00001**	**(101, 122)**
TYE7	**89°**	**0.79**	**(84, 95)**
NAM7	**93°**	**0.25**	**(80, 106)**

Table 4.11. Angles of the primary axis of variation estimated from principal component analysis on the mean level of expression against expression noise (both in z-score scale) for all genetic backgrounds. Statistical significance on whether the estimated angles were different from zero was estimated using bootstrap approach. P-values were estimated from 10000 bootstrap samples.

Genotype	Angle	P-value	95% CI
WT	90°	0.12	(1, 179)
M76	140°	< 0.00001	(129, 163)
TATA1	86.5°	0.0021	(83, 89)
M66	90°	0.64	(1, 179)
TATA2	50°	0.0001	(17, 86)
RAP1	90°	0.60	(1, 179)
ADE6	139°	< 0.00001	(126, 152)
TYE7	89°	0.79	(1, 179)
NAM7	93°	0.25	(1, 179)

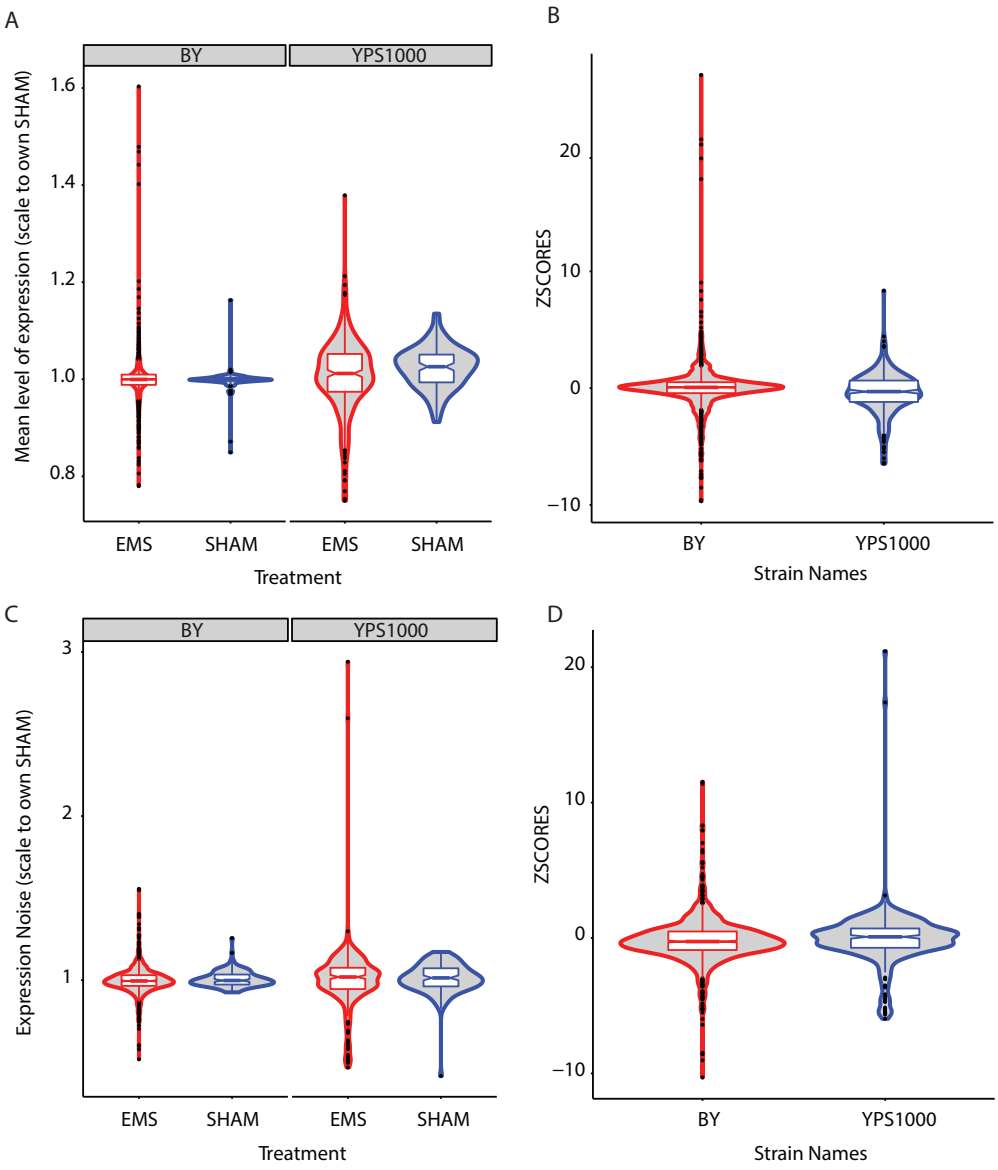

Figure 4.1. Mutational effects on both the mean level of expression and expression noise in different *Saccharomyces cerevisiae* strains. (A-B). Violin plots showing distributions of the mean level of expression for SHAM control population (blue) and EMS mutagenized population (red) using either percent of changes relative to the mean of SHAM **(A)** or z-scores **(B)**. **(C-D).** Violin plots showing distributions of expression noise for SHAM control population and EMS mutagenized population using either percent of changes relative to the mean of SHAM **(C)** or z-scores **(D)**.

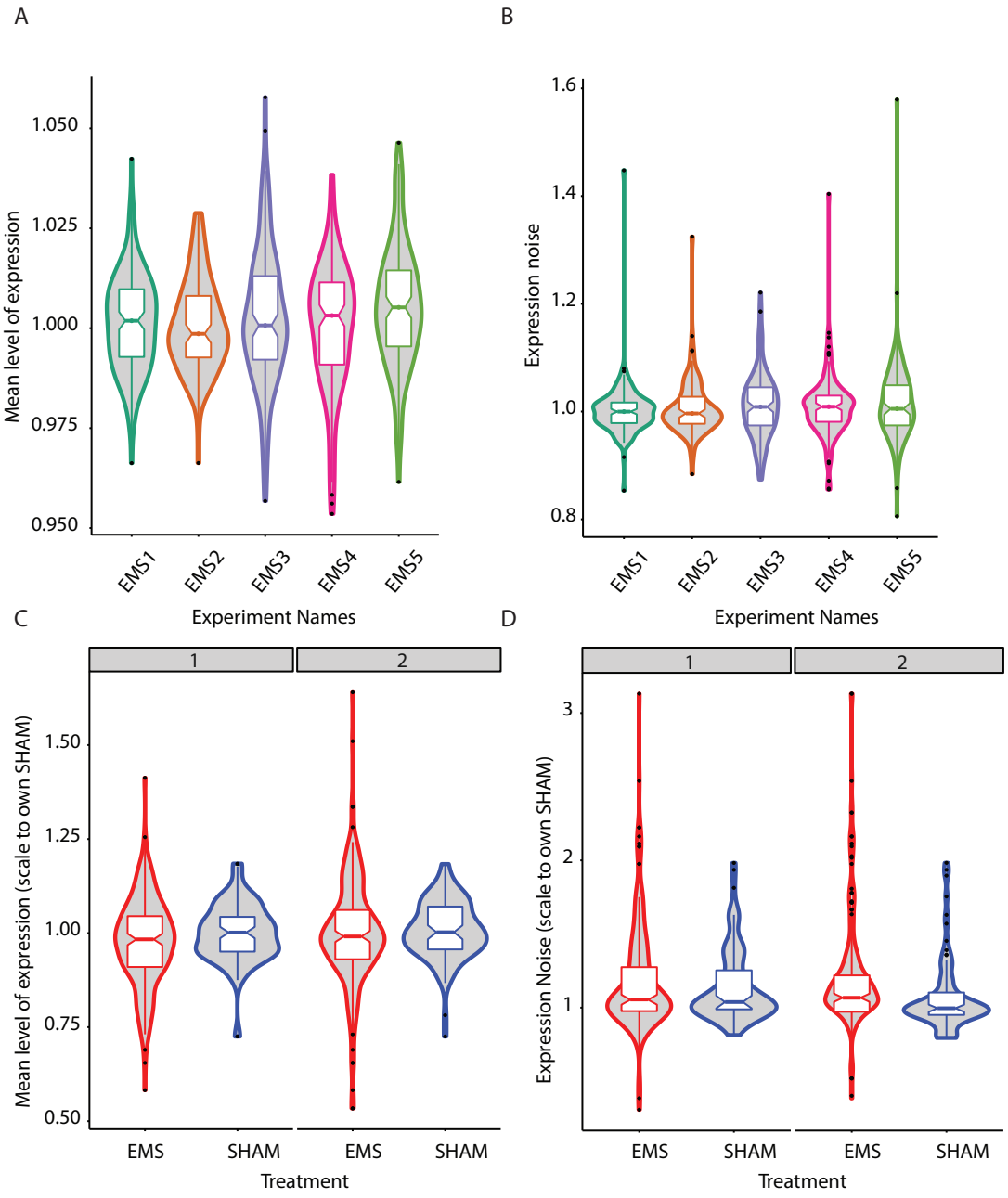

Figure 4.2. Quantifications of distributions on both the mean level of expression and expression noise were reproducible across different mutagenesis experiments. (A-B). Violin plots showing distributions of the mean level of expression (**A**) or expression noise (**B**) for SHAM control populations of BY strain in 5 different experiments. (**C-D**). Violin plots showing distributions of the mean level of expression (**C**) or expression noise (**D**) for both SHAM (blue) and EMS treated (red) populations in two independent mutagenesis experiments using M76 genotype.

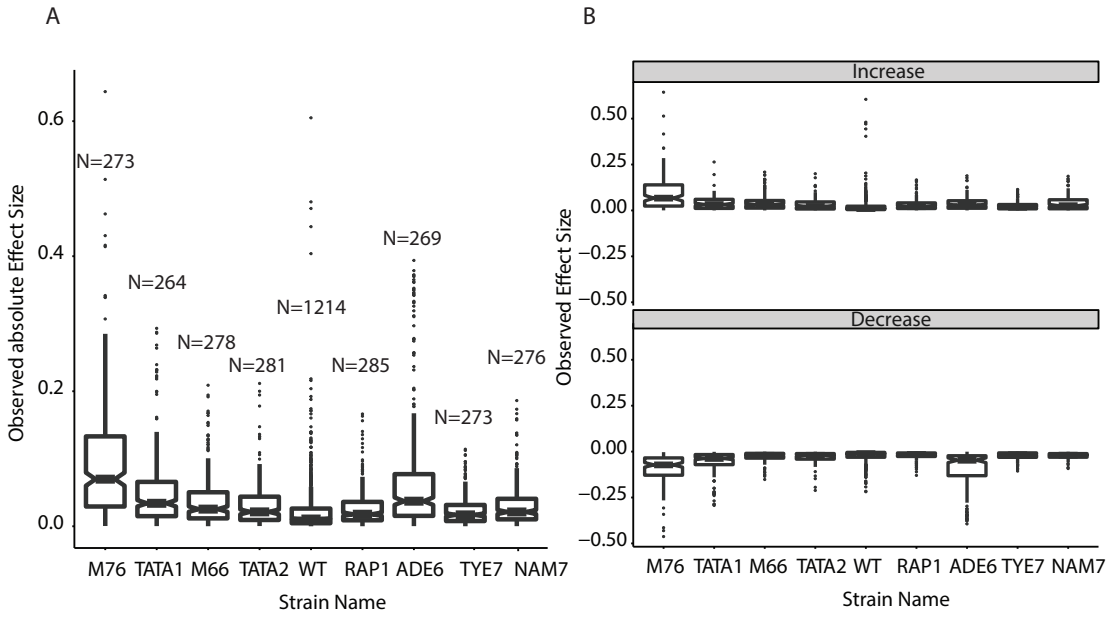

Figure 4.3. Comparisons of magnitude of mutational effects on the mean level of expression between BY strain and other starting genotypes. (A). Boxplots showing distributions of absolute value of magnitude of mutational effects for all genotypes. Numbers on top of each boxplot represent number of mutants for each genotype. **(B).** Boxplots showing distributions of mutational effects on the mean level of expression for mutants increasing (top) or decreasing (bottom) the mean level of expression.

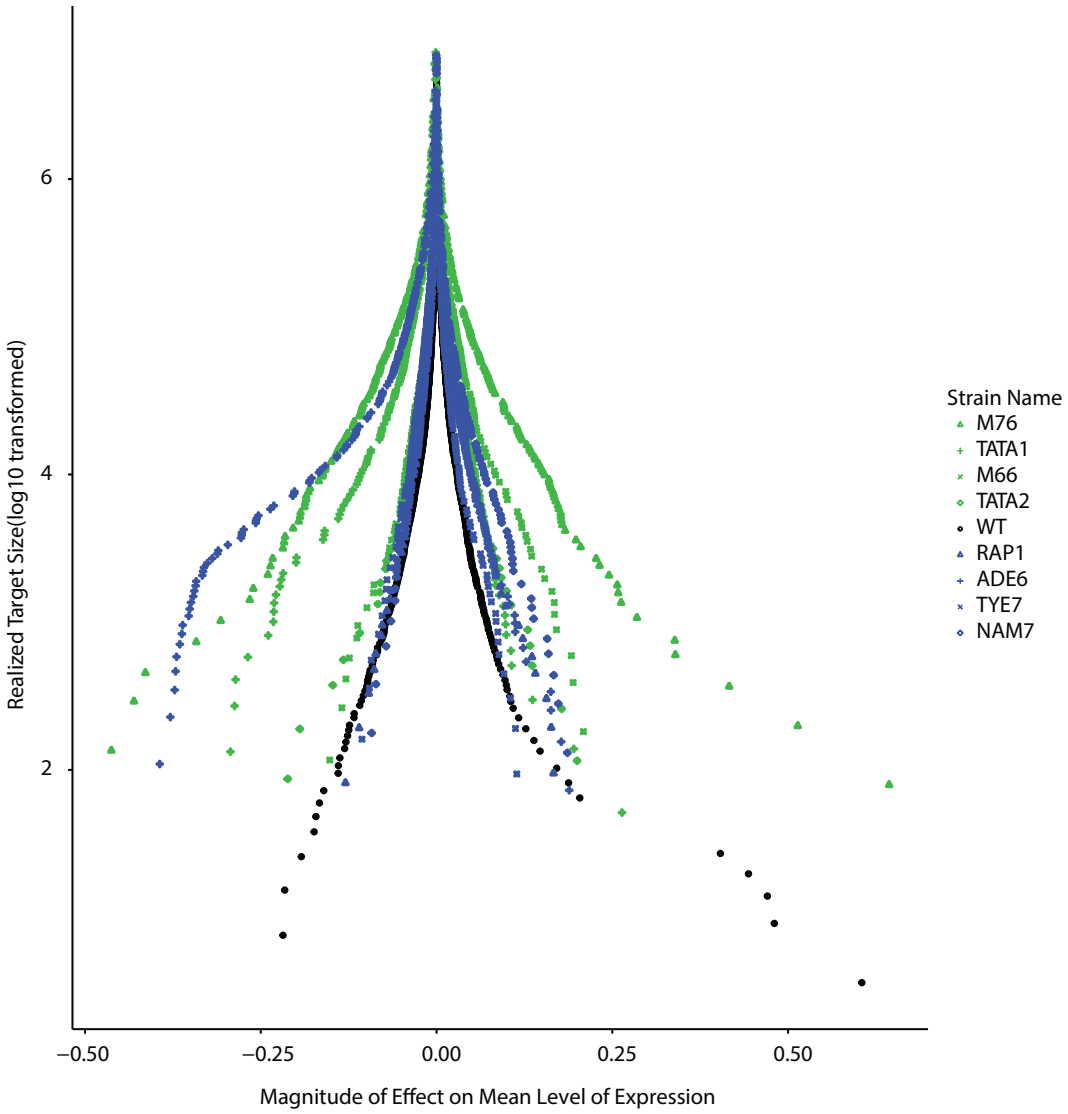

Figure 4.4. Estimations of mutational target size on the mean level of expression for different effect size cutoffs. For each genotype, each point in the figure represents that for a specific effect size cutoff (X-axis), number of nucleotides (Y-axis, log10 transformed) in the genome that when mutated would have effects on the mean level of expression equal or larger than the cutoff. Estimations for BY strain (black), 4 *cis* genetic variants (green) and 4 *trans* genetic variants (blue) are shown in the figure.

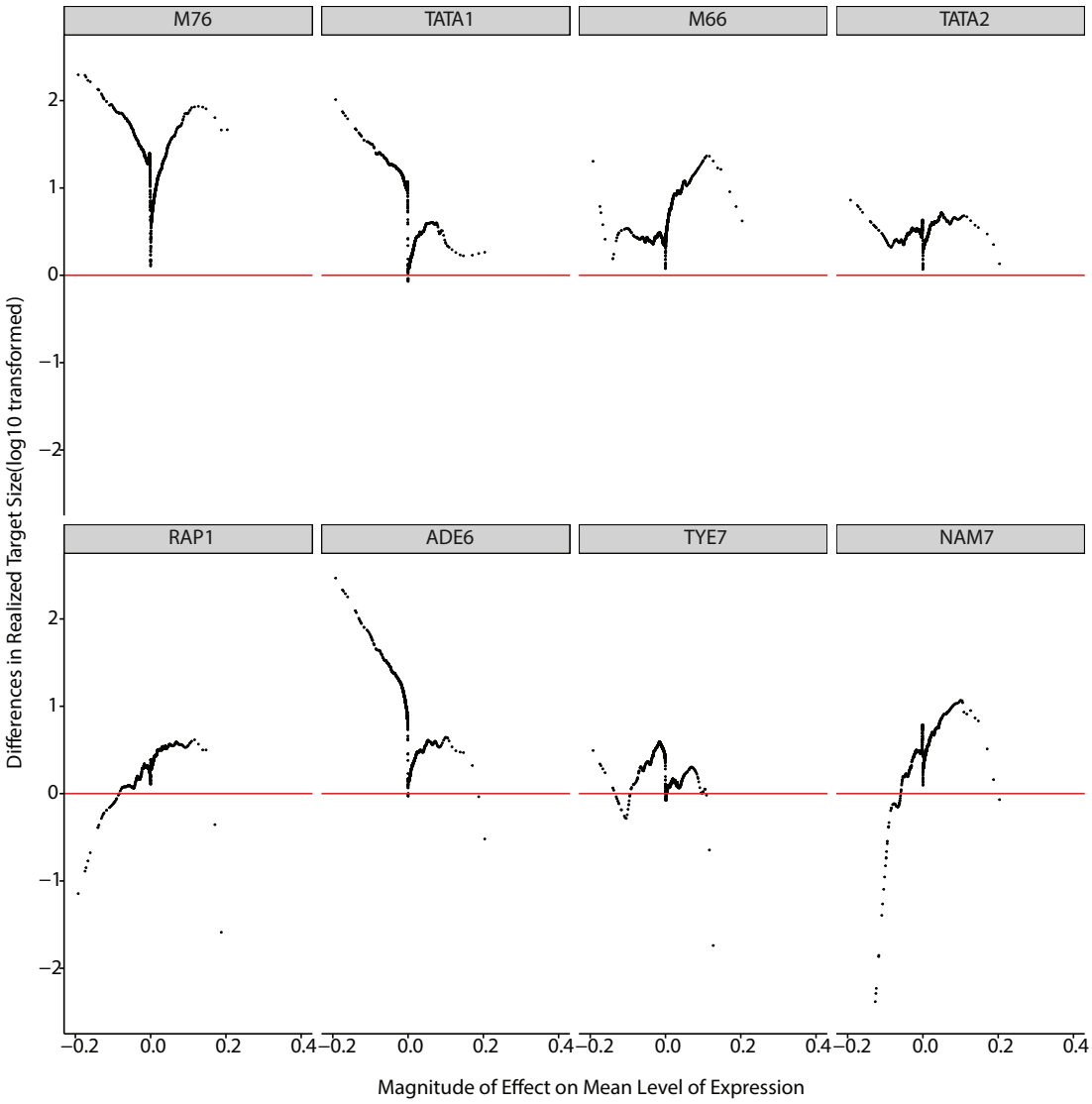

Figure 4.5. Differences in mutational target size for the mean level of expression between BY strain and all other genotypes for different effect size cutoffs. For each panel, each point represents that for a specific effect size cutoff (X-axis), differences in mutational target size (Y-axis, log10 transformed) estimated in Figure 4.4 between BY strain and the corresponding genotype in that panel.

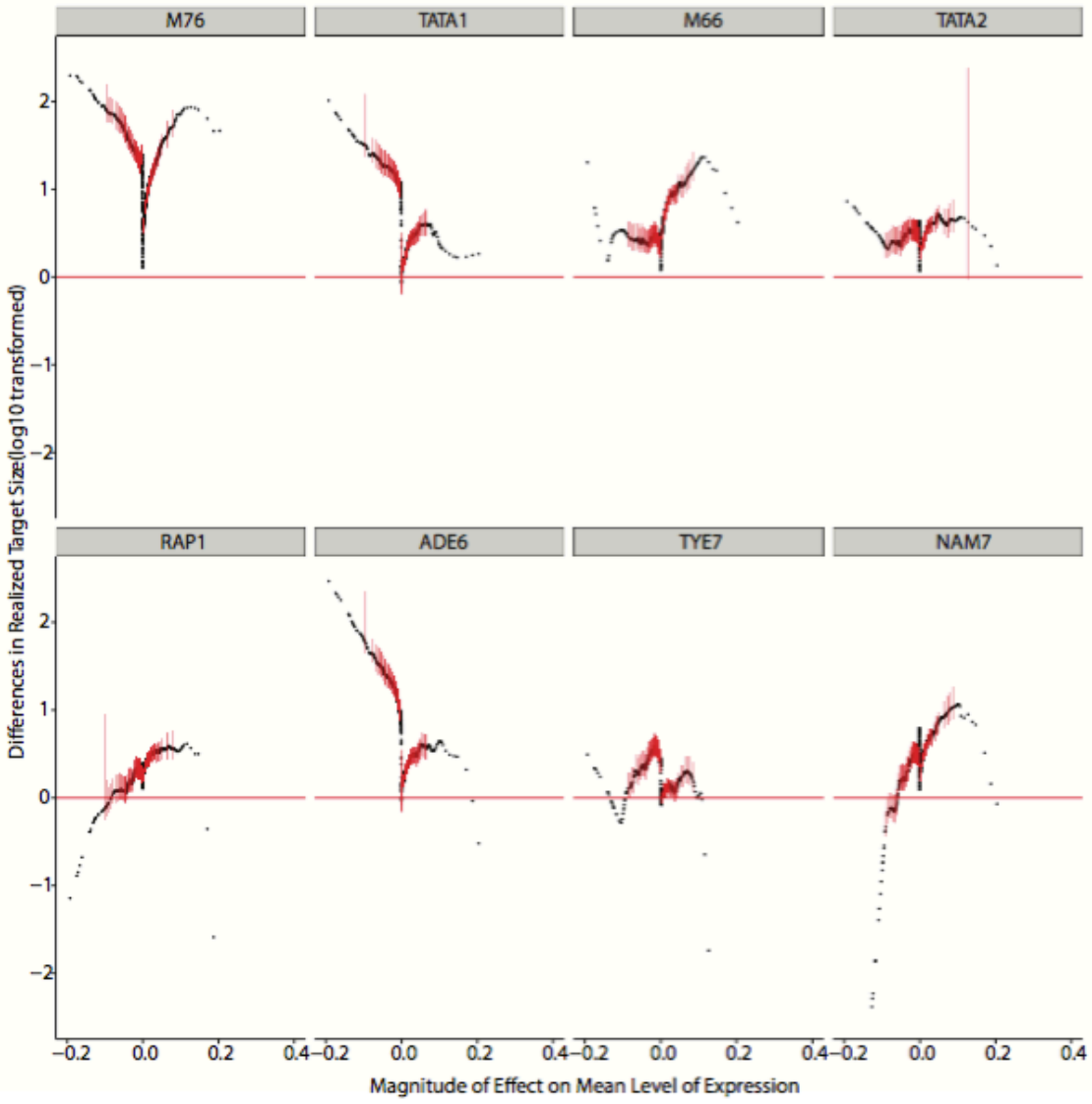

Figure 4.6. Differences in mutational target size for the mean level of expression between BY and other genotypes using random samples from the BY dataset. 200 random samples, each with similar number of mutants (~290) compared to all genotypes with genetic variants were drawn from mutants in BY background (~1210). Analysis in Figure 4.5 was repeated on each of the 200 random samples to calculate variations in estimating differences in mutational target size due to limited sample size. Red shades in each panel represent 95% confidence intervals estimated from 200 random samples. Overlapping between 95% CI and X-axis suggests that differences in mutational target size estimated in Figure 4.5 is not significant from zero if similar amount of mutants were collected for BY strain in multiple independent experiments.

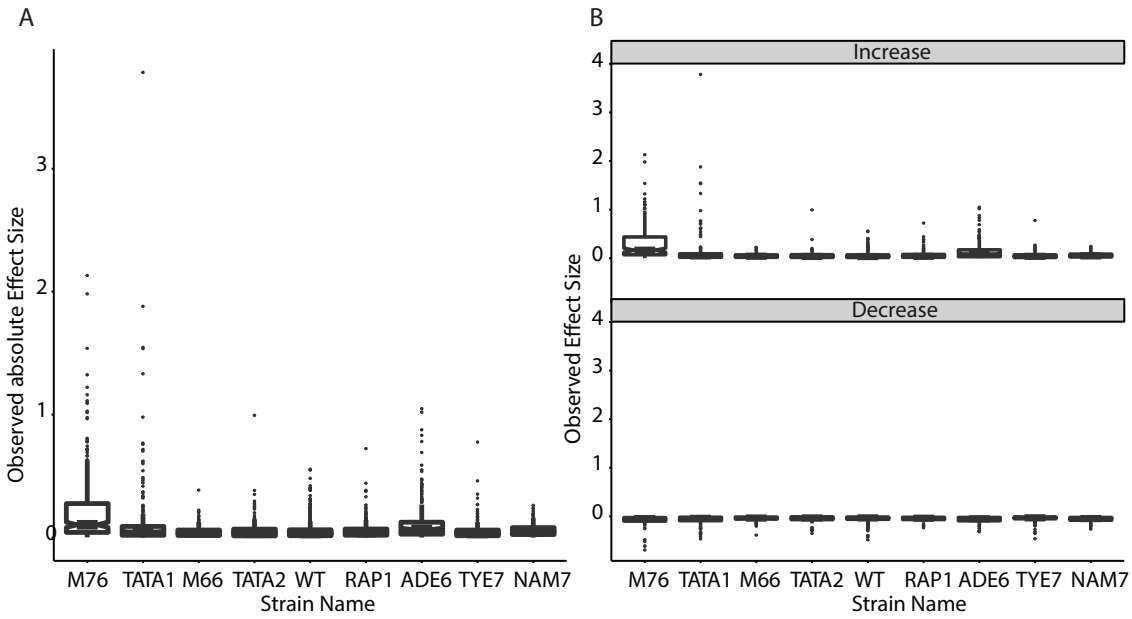

Figure 4.7. Comparisons of magnitude of mutational effects on expression noise between BY strain and other starting genotypes. (A). Boxplots showing distributions of absolute value of magnitude of mutational effects for all genotypes. Numbers on top of each boxplot represent number of mutants for each genotype. **(B).** Boxplots showing distributions of mutational effects on expression noise for mutants increasing (top) or decreasing (bottom) expression noise.

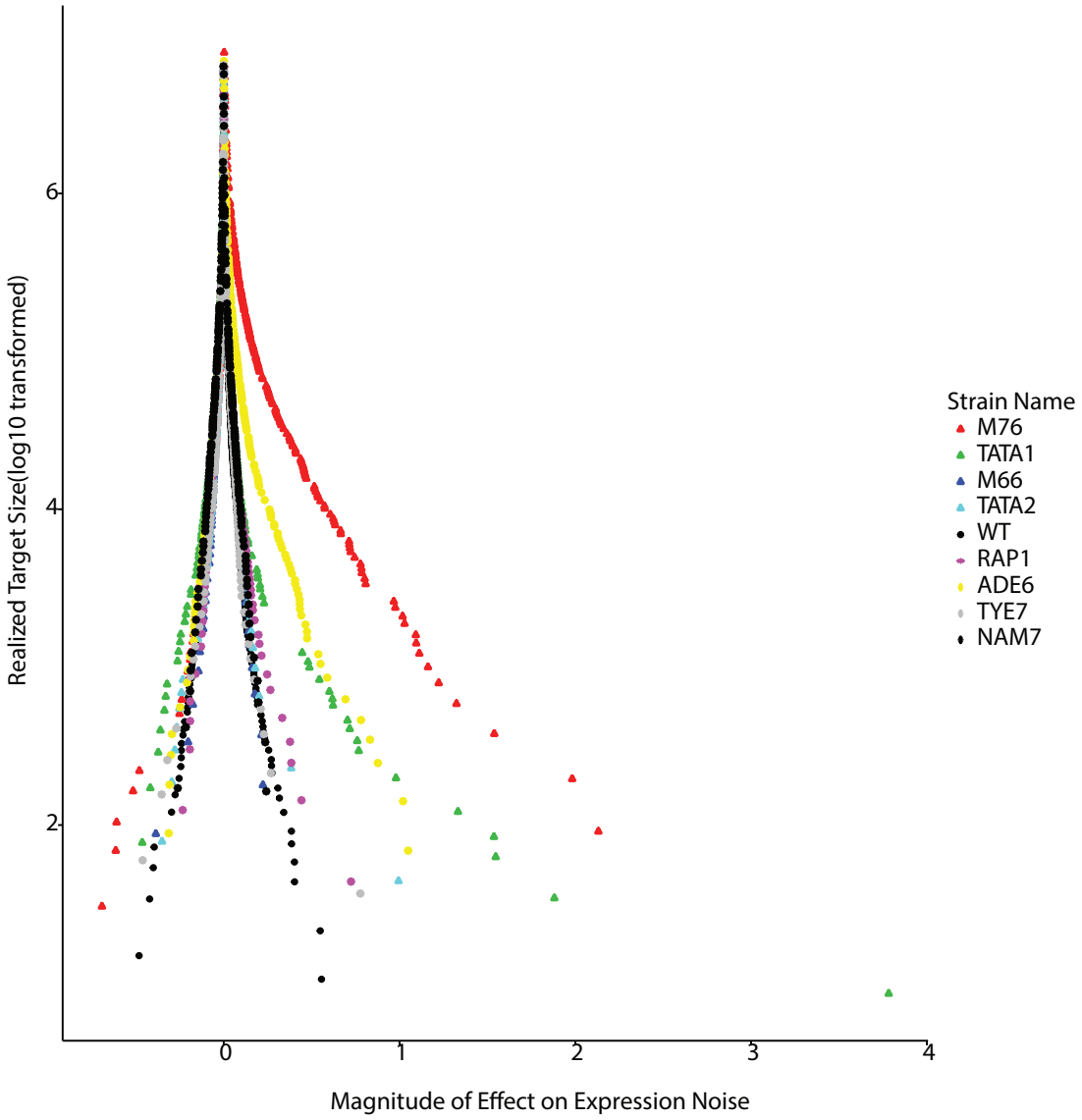

Figure 4.8. Estimations of mutational target size on expression noise for different effect size cutoffs. For each genotype, each point in the figure represents that for a specific effect size cutoff (X-axis), number of nucleotides (Y-axis, log10 transformed) in the genome that when mutated would have effects on expression noise equal or larger than the cutoff. Estimations for BY strain (black empty circle), 4 *cis* genetic variants (triangle) and 4 *trans* genetic variants (solid circle) are shown in the figure.

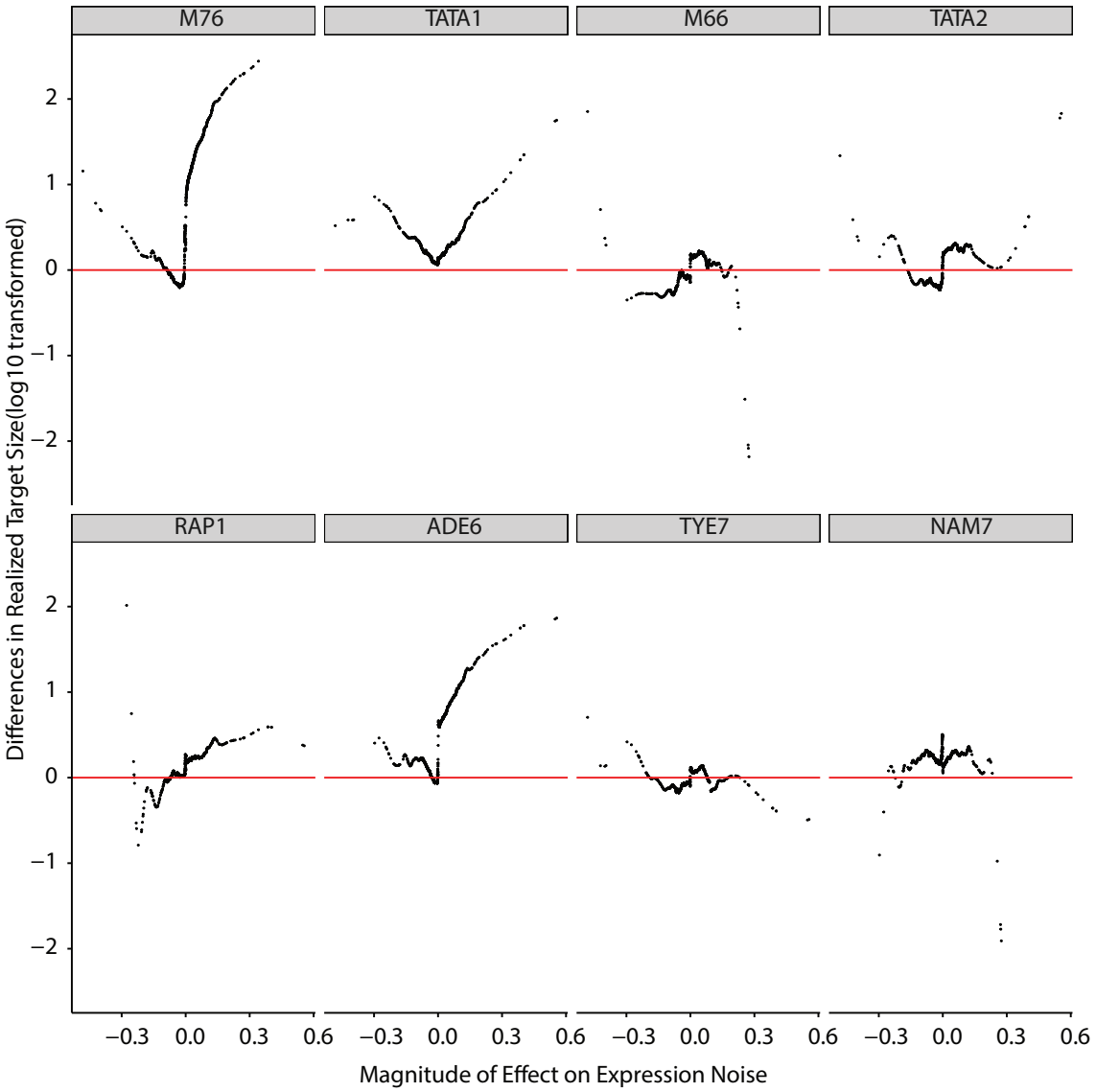

Figure 4.9. Differences in mutational target size on expression noise between BY strain and all other genotypes for different effect size cutoffs. For each panel, each point represents that for a specific effect size cutoff (X-axis), differences in mutational target size (Y-axis, log10 transformed) estimated in Figure 4.8 between BY strain and the corresponding genotype in that panel.

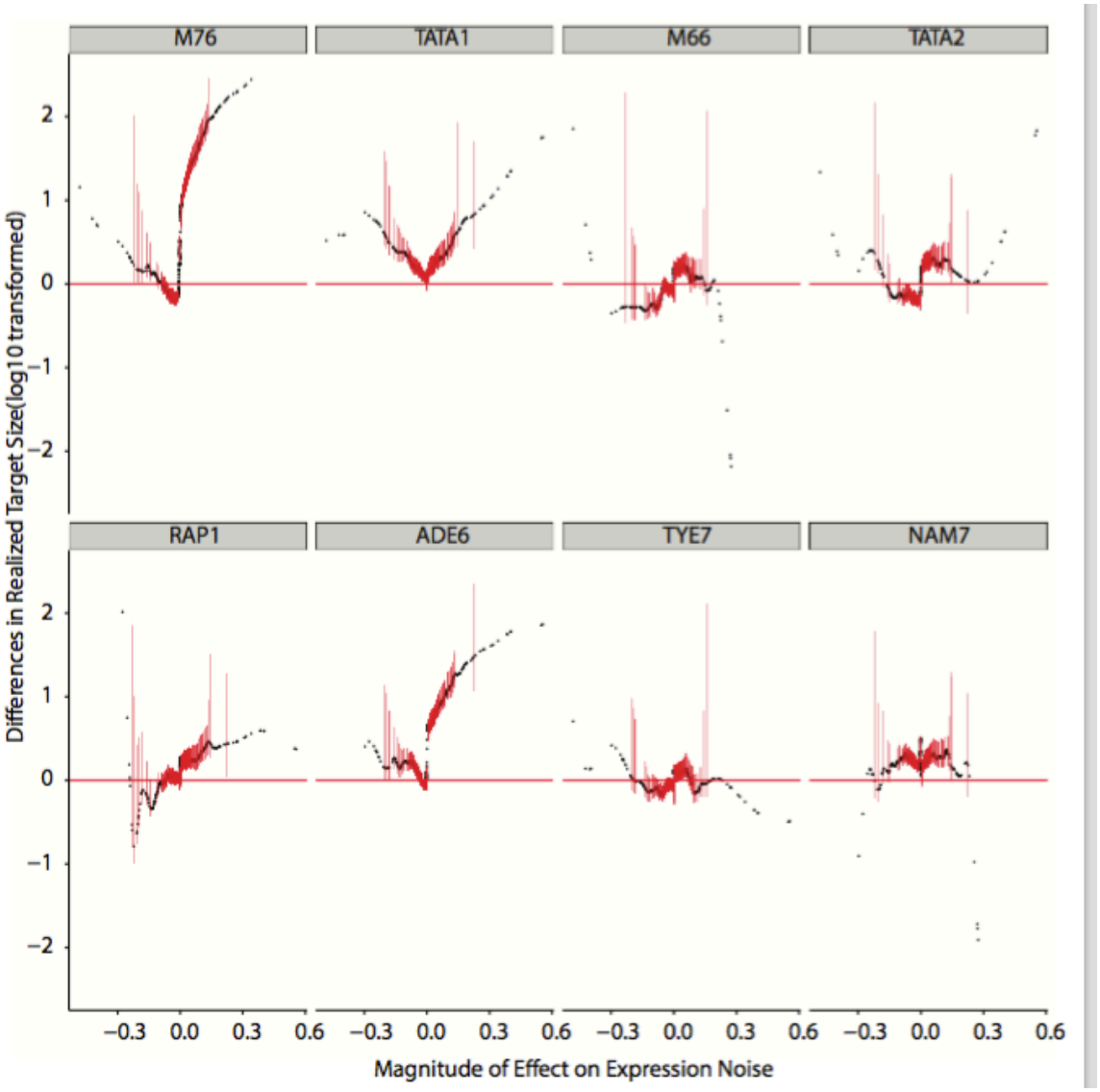

Figure 4.10. Differences in mutational target size for expression noise between BY and other genotypes using random samples from the BY dataset. 200 random samples, each with similar number of mutants (~290) compared to all genotypes with genetic variants were drawn from mutants in BY background (~1210). Analysis in Figure 4.9 was repeated on each of the 200 random samples to calculate variations in estimating differences in mutational target size due to limited sample size. Red shades in each panel represent 95% confidence intervals estimated from 200 random samples. Overlapping between 95% CI and X-axis suggests that differences in mutational target size estimated in Figure 4.5 is not significant from zero if similar amount of mutants were collected for BY strain in multiple independent experiments.

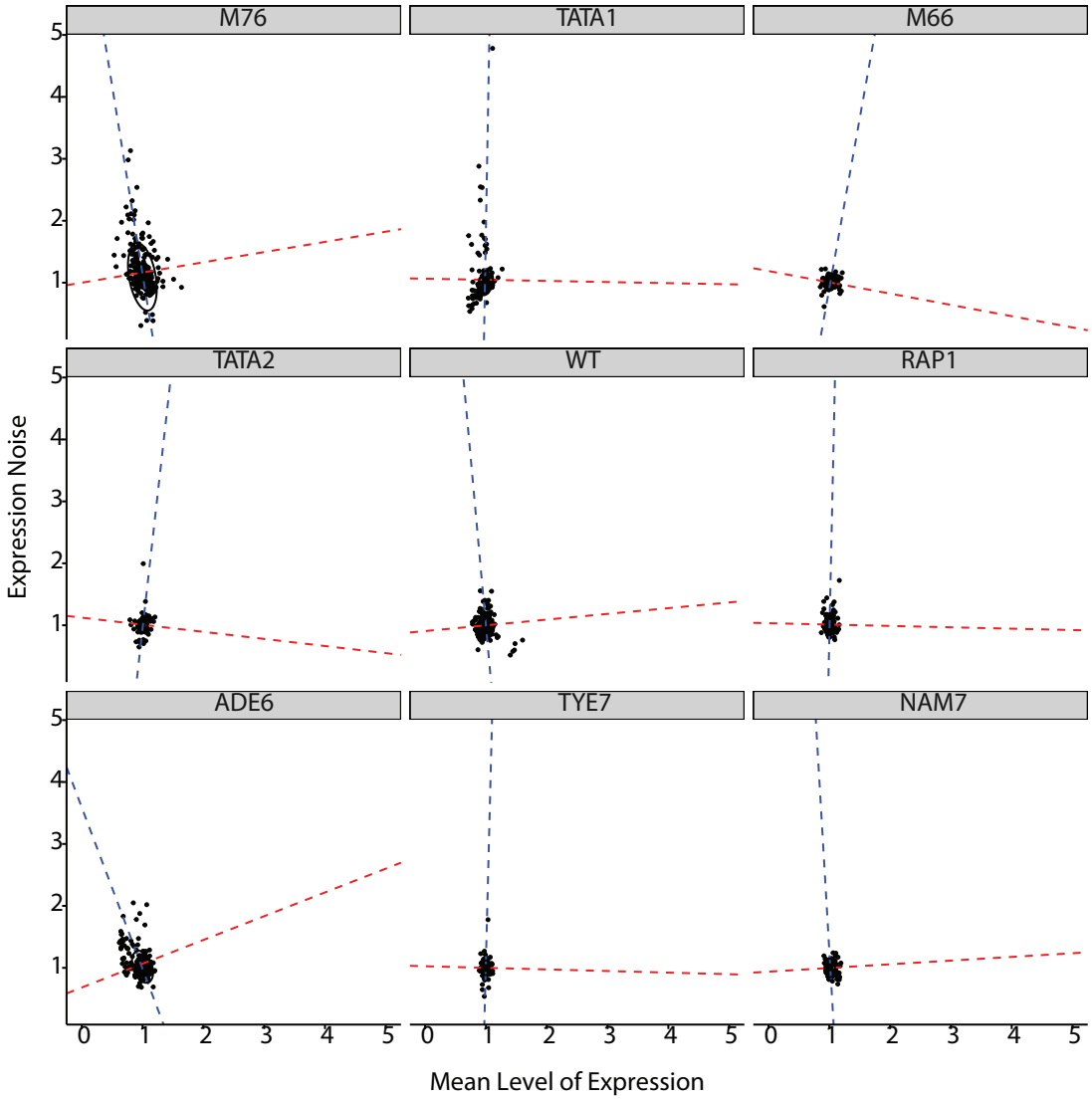

Figure 4.11. Relationship between the mean level of expression and expression noise in different starting genetic backgrounds. In each panel, the mean level of expression (X-axis) is plotted against expression noise (Y-axis) for EMS treated populations (using percent change relative to average of SHAM populations). Blue dashed line represents direction of primary axis of variation from Principal Component Analysis (PCA). Red dashed line represents direction of secondary axis of variation.

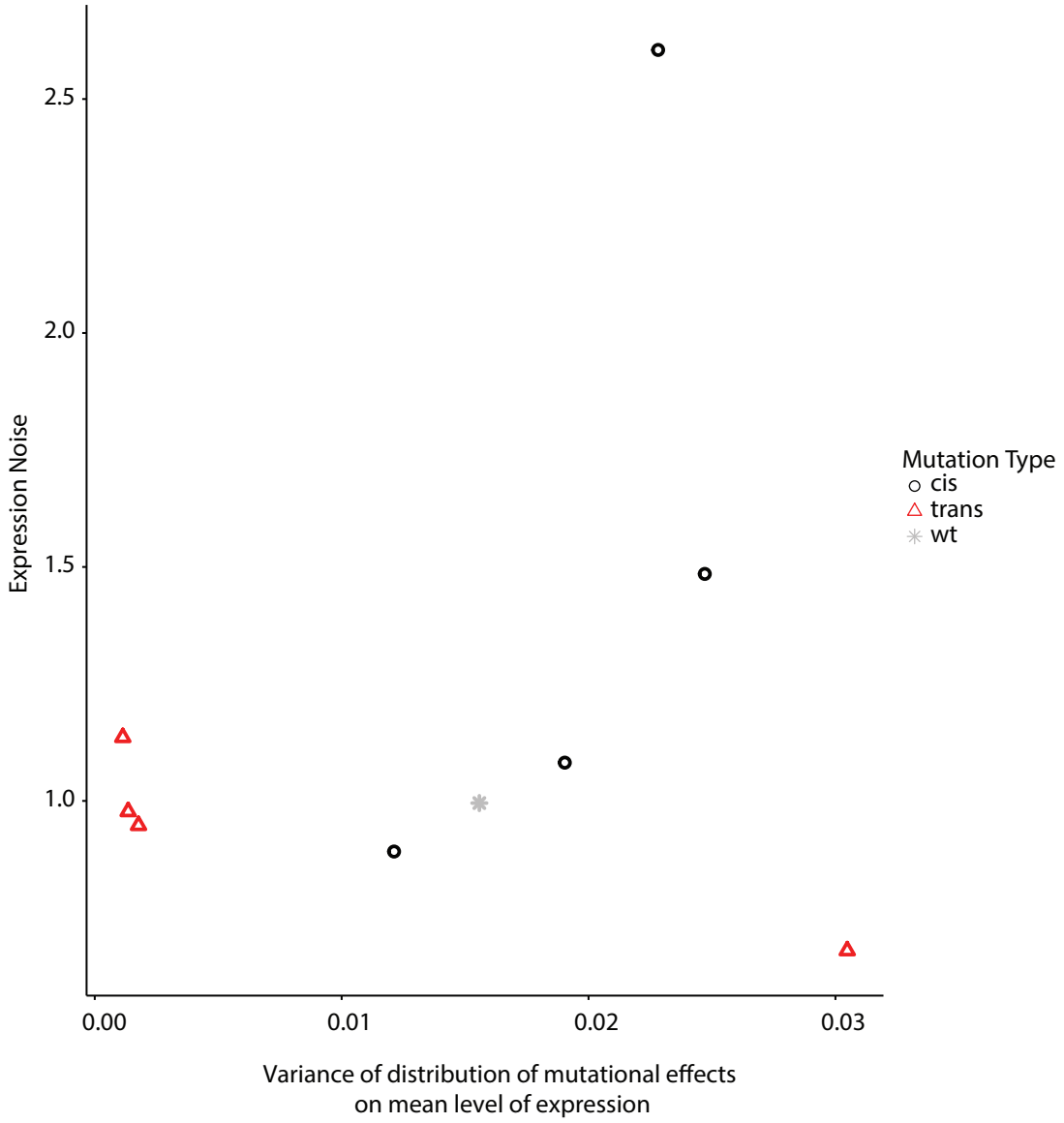

Figure 4.12. Relationship between expression noise and variation of mutational effects on the mean level of expression across different genotypes. The expression noise estimated as mean noise from SHAM population for each genotype (Y-axis) is plotted against variance of distribution of mutational effects on the mean level of expression (X-axis).

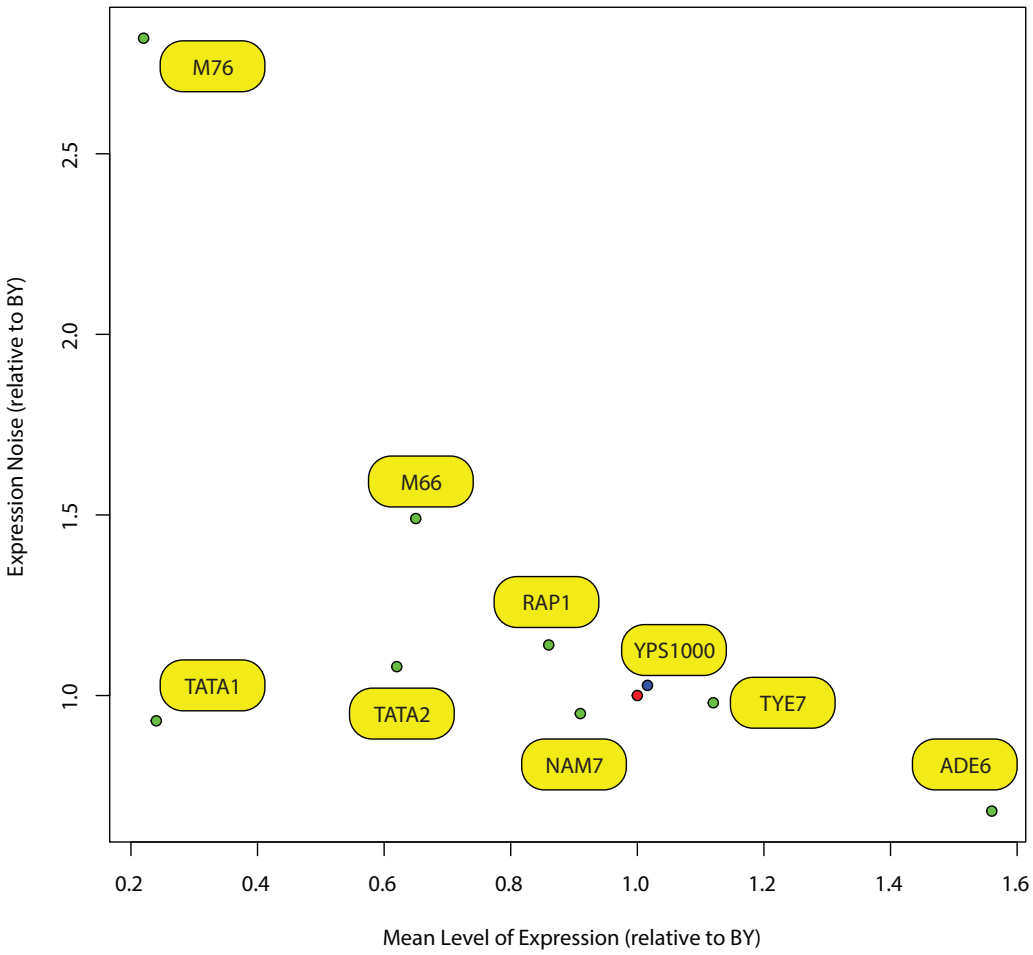

Figure 4.13. Level of average expression and expression noise relative to BY strain for all other genetic backgrounds. Expression noise (Y-axis) is plotted against the mean level of expression (X-axis) for BY (red), YPS1000 (blue) and all other genotypes carrying genetic variants (green).

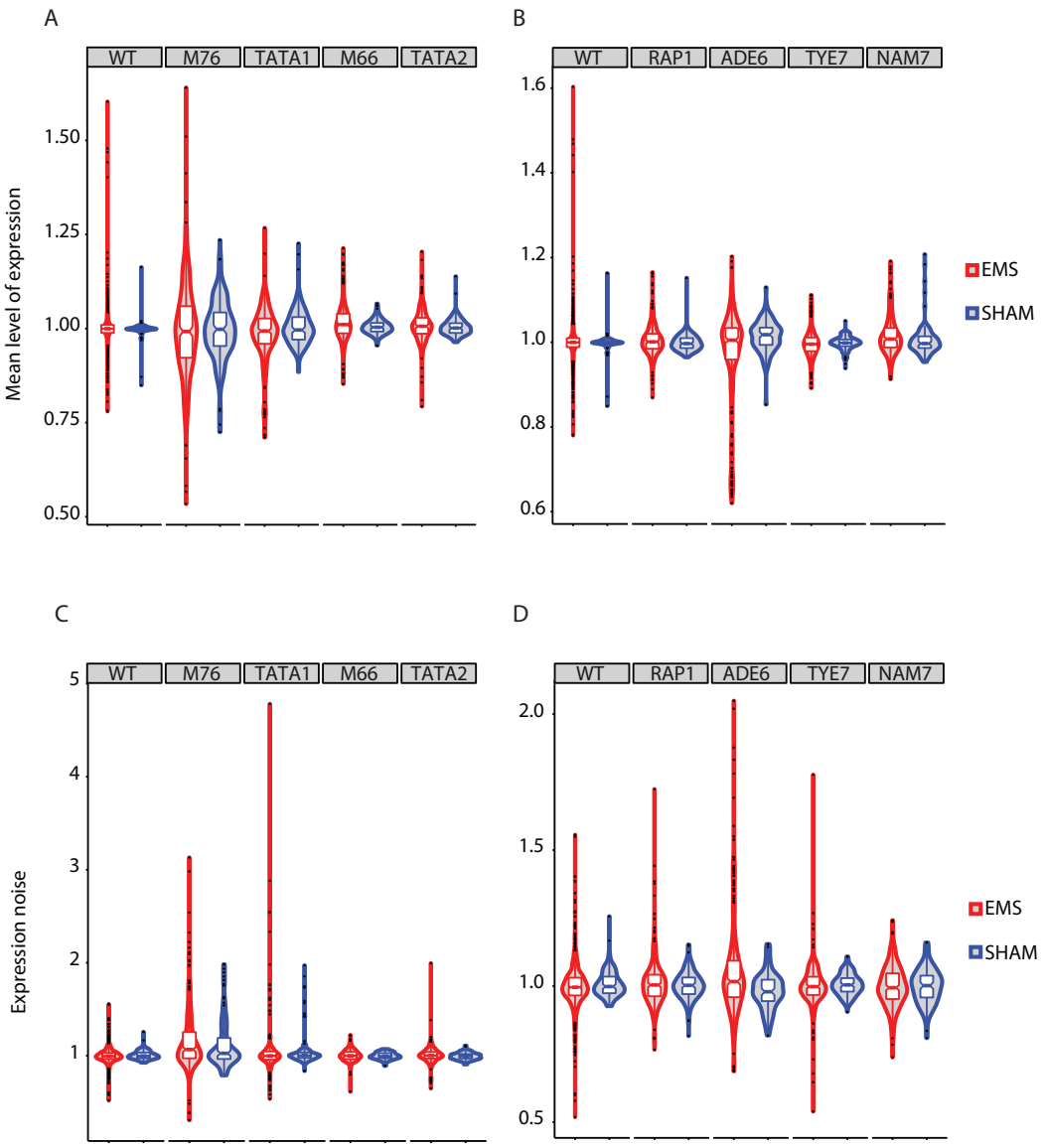

Figure 4.14. Distributions of mutational effects on the mean level of expression and expression noise for genotypes with genetic variants. Violin plots showing distributions of the mean level of expression (**A-B**) and expression noise (**C-D**) for both SHAM populations (blue) and EMS treated populations (red). Scale on Y-axis is percent change relative to mean of SHAM populations for each genotype respectively.

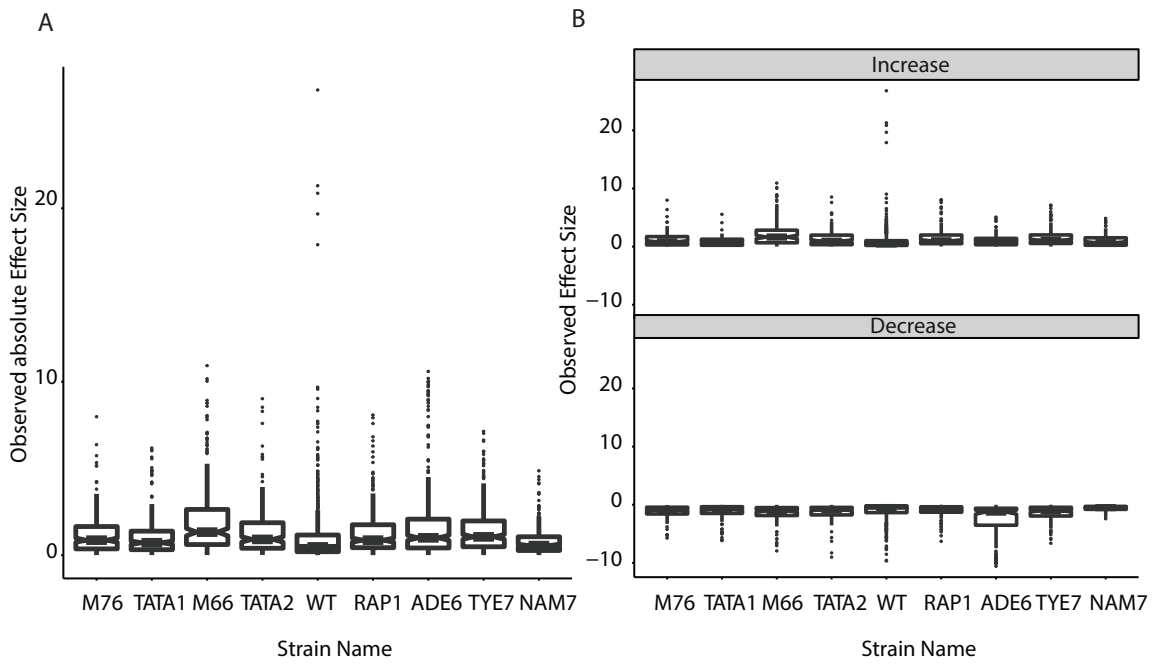

Figure 4.15. Comparisons of magnitude of mutational effects on the mean level of expression between BY strain and other starting genotypes using z-score. (A). Boxplots showing distributions of absolute value of magnitude of mutational effects (in z-score) for all genotypes. Numbers on top of each boxplot represent number of mutants for each genotype. **(B).** Boxplots showing distributions of mutational effects on the mean level of expression (in z-score) for mutants increasing (top) or decreasing (bottom) the mean level of expression.

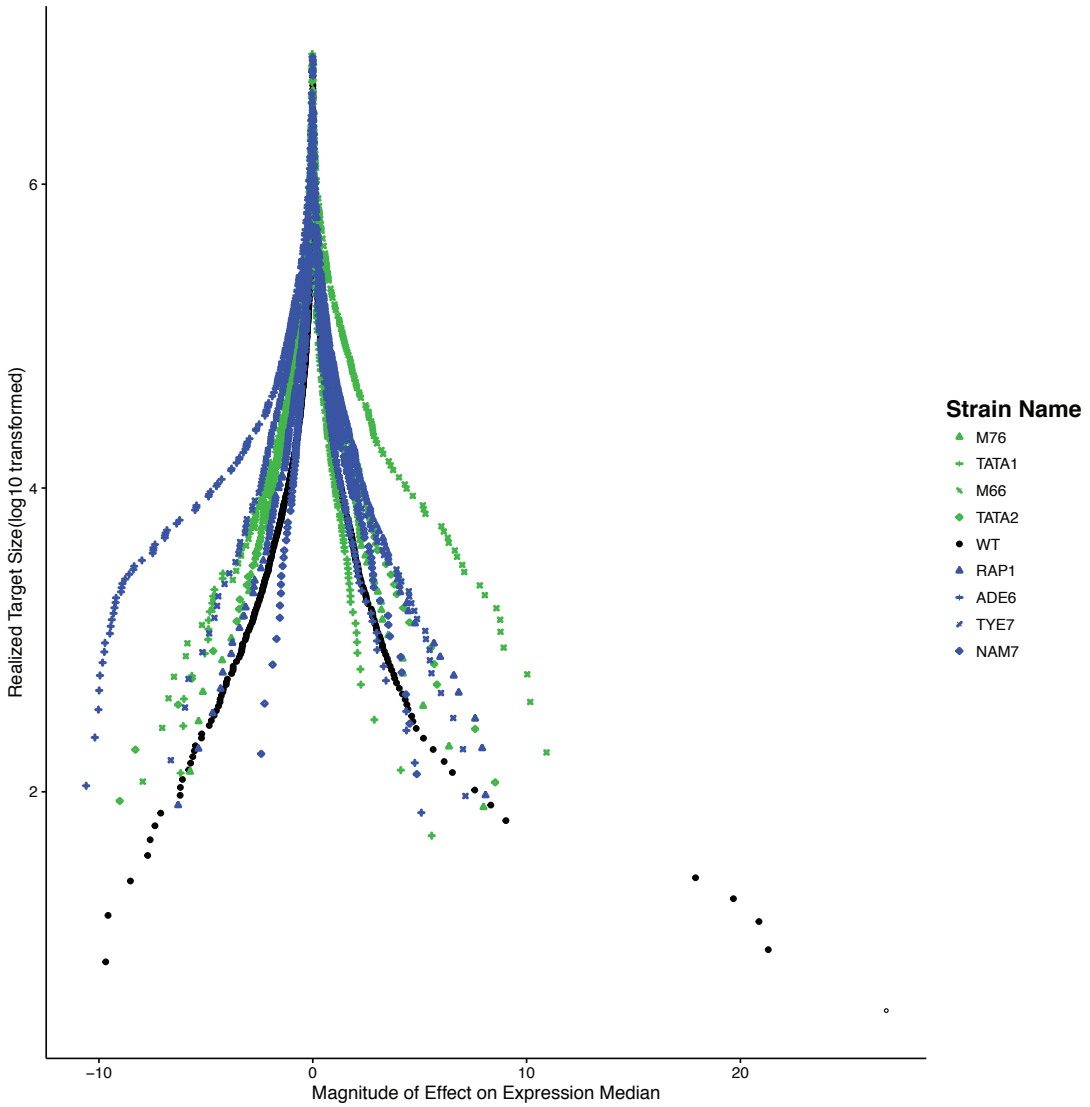

Figure 4.16. Estimations of mutational target size on the mean level of expression for different effect size cutoffs using z-score. For each genotype, each point in the figure represents that for a specific effect size cutoff (in z-score) (X-axis), number of nucleotides (Y-axis, log10 transformed) in the genome that when mutated would have effects on the mean level of expression equal or larger than the cutoff. Estimations for BY strain (black), 4 *cis* genetic variants (green) and 4 *trans* genetic variants (blue) are shown in the figure.

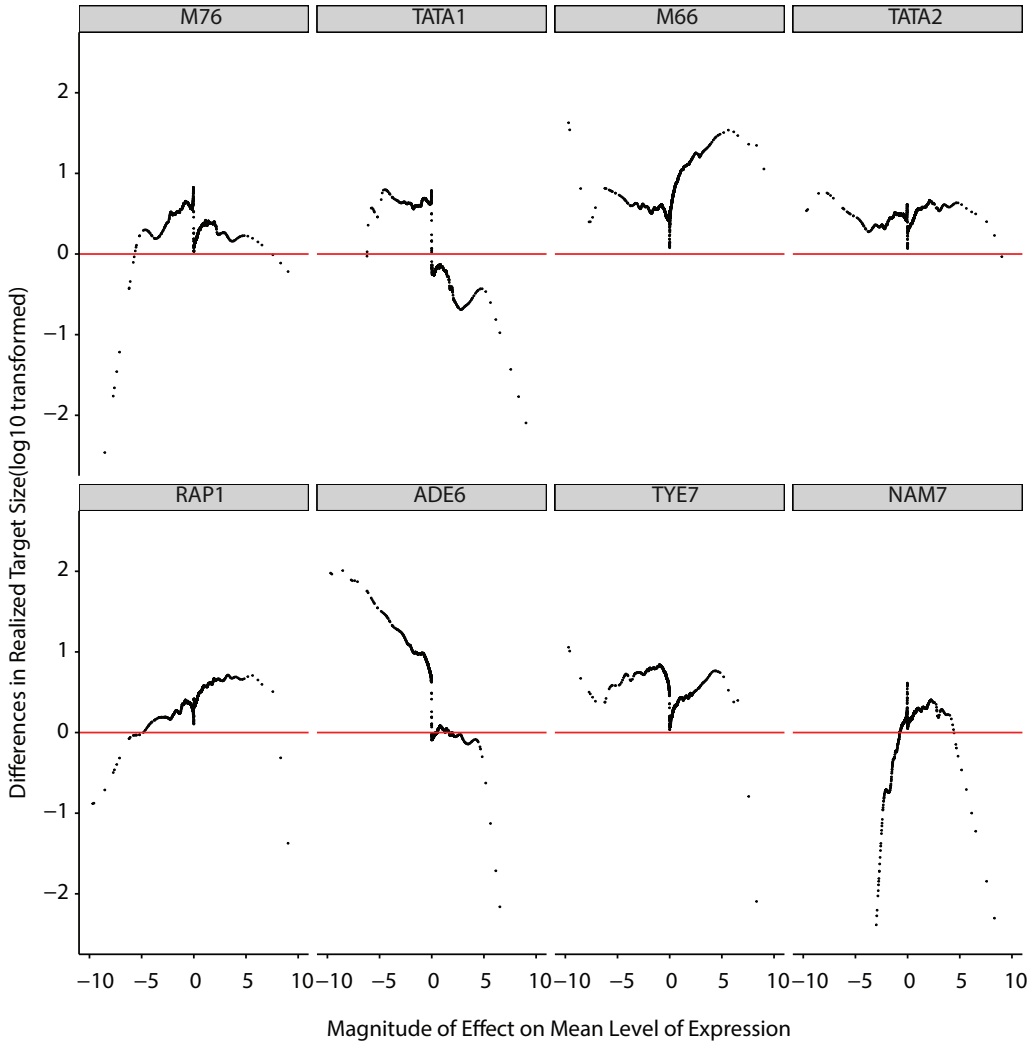

Figure 4.17. Differences in mutational target size between BY strain and all other genotypes for the mean level of expression on different effect size cutoffs in z-score. For each panel, each point represents that for a specific effect size cutoff (in z-score) (X-axis), differences in mutational target size (Y-axis, log10 transformed) estimated in Figure 4.16 between BY strain and the corresponding genotype in that panel.

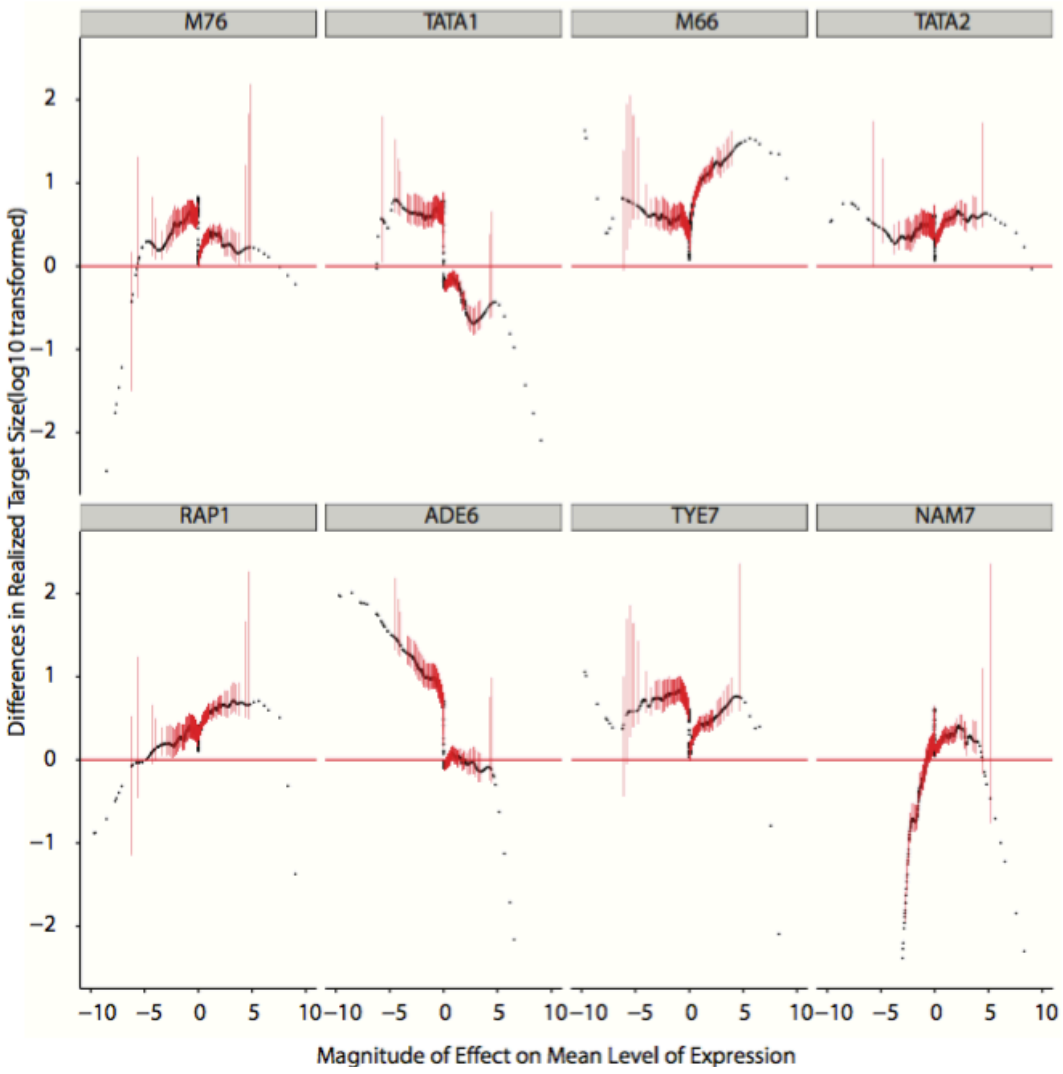

Figure 4.18. Differences in mutational target size for the mean level of expression between BY and other genotypes using random samples from BY dataset using z-score. 200 random samples, each with similar number of mutants (~290) compared to all genotypes with genetic variants were drawn from mutants in BY background (~1210). Analysis in Figure 4.17 was repeated on each of the 200 random samples to calculate variations in estimating differences in mutational target size due to limited sample size. Red shades in each panel represent 95% confidence intervals estimated from 200 random samples. Overlapping between 95% CI and X-axis suggests that differences in mutational target size estimated in Figure 4.5 is not significant from zero if similar amount of mutants were collected for BY strain in multiple independent experiments.

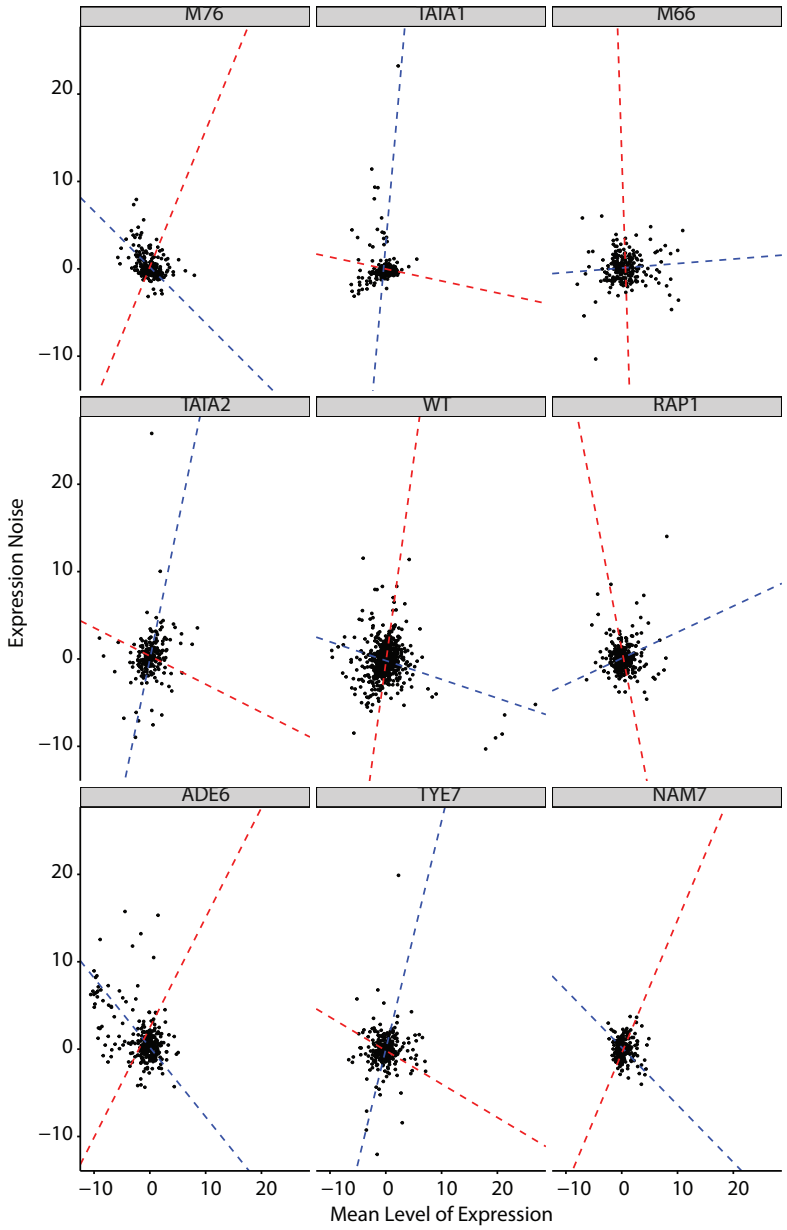

Figure 4.19. Relationship between the mean level of expression and expression noise in different starting genetic backgrounds using z-score. In each panel, the mean level of expression (in z-score) (X-axis) is plotted against expression noise (in z-score) (Y-axis) for EMS treated populations (using percent change relative to average of SHAM populations). Blue dashed line represents direction of primary axis of variation from Principal Component Analysis (PCA). Red dashed line represents direction of secondary axis of variation.

Chapter V

Concluding remarks

The variations in the transcriptional regulation is a major contributor to the phenotypic evolution (Wray 2007; Carroll 2008; Stern and Orgogozo 2008). In the recent decades, studies on the transcriptional regulatory network have generated many new insights on the molecular mechanisms of the transcriptional regulation. Those studies suggest that the organizations of regulatory interactions among transcription factors and their target genes influence the transcriptional outcomes (Davidson et al. 2002; Alon 2007). However, whether and how the structural properties of the transcriptional regulatory network affect the evolution of gene expression is less clear. In this thesis, I studied whether and how two connective properties of the transcriptional regulatory network, the in-degree and the out-degree, were associated with the differences in gene expression over evolutionary time (Chapter 2 and Chapter 3). I also studied how the genetic changes that directly or indirectly disrupt the regulation of the transcription of the gene *TDH3* affect mutational effects on both the mean level of expression and the expression noise (Chapter 4). Below I will discuss some implications from each study as well as some potential directions to better understand those questions.

The impact of the regulatory network on the evolution of gene expression depends on the biological context

One of the central questions in the evolutionary biology is whether the genetic basis of the phenotypic evolution is predictable, and if so, what biological factors are important in determining the evolutionary consequences. Specific to the evolution of the gene expression, the observations that the genetic basis for the phenotypic innovation in parallel evolution are similar across different species (Stern and Orgogozo 2008; Stern and Orgogozo 2009) motivate the hypothesis that there exist some biological properties for different genes, although currently unclear to evolutionary biologists, that impact the trajectory of the evolution of expression. I discussed several properties in Chapter 2-5, including but not limited to the effect of pleiotropy, the mutational target size in the genome and the amount of genetic interactions among genetic variations described by the level of epistasis. The transcriptional regulatory network, as a systematic representation of the regulatory interactions, can be useful for understanding the problem of the predictability of the evolution of gene expression in at least two perspectives..

First, many important properties that have been predicted to influence the evolution of gene expression can be examined in the context of the regulatory network. For example, in Chapter 2 and 3, I showed that the number of regulators (or in-degree) for a gene, which is a commonly used metric in describing the connectivity of a node in the network literatures (Newman 2005), can be used to validate the predictions from two conflicting evolutionary models. More specifically, the in-degree is predicted to be associated with either the robustness to the variations in gene expression (MacNeil and Walhout 2011) or the amount of genetic variations available to change the expression

level (Featherstone and Broadie 2002; Landry et al. 2007). My results suggest that both models can be informative in understanding the evolution of gene expression. However, the dominant force that might impact the evolution of gene expression depends on the specific systems used. This is a good example showing the power of the regulatory network in testing whether the biological properties predicted to be important in the evolution of gene expression from theoretical studies have the expected association with the observed patterns of the differences in gene expression..

Second, the regulatory network can generate the hypotheses that are difficult to be realized if one only focuses on studying the functions of individual genes. The *shavenbaby* story (McGregor et al. 2007) illustrates the idea that the genes in the central position (or network hub) in the genetic networks are the hotspots for phenotypic innovations. In contrary, multiple studies suggest that the expression levels of the network hubs in the genetic interaction networks or the regulatory networks are more conserved than other genes with less interactive partners (Lu et al. 2007; Goymer 2008). Although contradictory in their conclusions, all studies imply that the abstract concept of the network hub is an important property that is illustrative in answering the question of the predictability of the evolution of gene expression. Although my results do not directly generate hypotheses on new biological properties, the analysis in the *Drosophila* species suggests that the presence of redundant transcription factors for the regulation of gene expression, which is the genetic basis of the "hierarchical" organization of the regulatory interactions (Davidson et al. 2002), might provide the robustness to the variations in gene expression. This result is consistent with the findings that the master regulators in animal development tend to have more regulators than other genes

(Borneman et al. 2006; Vermeirssen et al. 2007). With deeper analysis on the structural properties of the regulatory network, as well as the introduction of new concepts from other field of science, more hypotheses could be generated on understanding the predictability of the evolution in gene expression..

Although the transcriptional regulatory network provides an interesting angle for studying the evolution of gene expression, there are multiple issues when we try to answer evolutionary questions in the context of the regulatory network. For example, although the regulatory network provides information in examining the roles of important genetic properties in the evolution of gene expression, the full conceptual framework associated with those properties might not be captured by the structural metrics within the network. In Chapter 2 and 3, the rational of the hypothesis that out-degree is associated with the evolution of gene expression is from the theory of pleiotropy. The lack of statistically significant association between differences in gene expression and the number of targets (out-degree) for a transcription factor in my study might reflect the fact that the number of targets is not a good measure of the level of pleiotropy. However, my analysis still provides a unique angle in understanding this evolutionary model. One should be cautious in drawing strong conclusions in similar genomic studies, since the analysis might not capture the true biological factors under consideration.

Second, since the observed evolutionary consequences are shaped by different evolutionary forces, the type of association between the structural properties and differences in gene expression might vary in different biological systems, which obscure our understanding of the roles of those properties in constraining the evolutionary process. For example, in Stern and Orgogozo. (2009), the authors pointed out that the

strength of selection and the population history, which are non-genetic factors, influence the evolution of gene expression. In my study, the inconsistency of conclusions on the in-degree between fly and yeast species could be due to differences in the level of species divergence and the population sizes in nature. This result highlights the necessity of having a more complete description of biological forces that could impact the evolution of gene expression. On the other hand, my results suggest that an accurate examination of how the structural properties influence the evolutionary trajectory requires a more controlled artificial system, in which only one or a few biological forces that are predicted to be important in the evolution of gene expression can vary while holding other factors constant. As a future extension, one way to check whether the in-degree or the out-degree affect the evolution of gene expression is through combining systems developed from the field of synthetic biology and the experimental evolution approaches. For example, an artificial genetic network could be designed and inserted into a single cell organism in which varying number of regulators are connected to a reporter construct expressing products important for survival of the organism in specific experimental conditions (Peisajovich 2012). Artificial selection pressure based on the reporter construct could then be applied to the engineered population, and the survival genotypes of the experimental evolution could be checked through identifying mutations that affect the fitness of individuals within the population. By using this experimental approach, we could show whether our predictions are close to truth in real biological systems.

Finally, any attempts to reconstruct the genome-wide regulatory network suffer from the inaccuracy in inferring the individual regulatory interaction. This problem becomes even more serious for comparative studies using multiple species. In my

analysis, although I provided several explanations for the inconsistency on the conclusions for the in-degree between intra-species comparison and inter-species comparisons, I cannot rule out the possibility that the structure of the regulatory network has changed over time. Fortunately, more efforts have been put to collect genomic data on species other than the widely used model organisms. For example, genomes of more natural strains and species in *Saccharomyces* clades have been sequenced (Kellis et al. 2003; Carreto et al. 2008). Also, more efforts have been taken to collect data on regulatory interactions in different *Drosophila* species and *Caenorhabditis* species through modENCODE projects (Boley et al. 2014). Thus, a future extension on my work would be to combine species-specific genomic datasets to reconstruct the regulatory network for different species independently. By this approach, we can not only have more accurate information on the structural properties in comparative studies, but also have the chance to examine the roles of structural changes of the regulatory network during the evolution of gene expression.

Taken together, my analysis in Chapter 2 and 3 illustrate both the strength and the drawbacks of using the regulatory network in understanding the evolution of gene expression. It should be noted that the benefits of using the regulatory network in evolutionary studies does not dishonor the more traditional candidate gene approaches. In fact, all the insights that motivate the studies on the network are from old-school experimental systems, and they are still the most important basis in the evolutionary studies. However, the success of using the regulatory network in understanding the evolution of gene expression suggests that more efforts should be taken to combine insights generated from network biology and the conceptual frameworks from the

traditional evolutionary models, so that we could have a more systematic view of figuring out the problem of predictability of phenotypic evolution..

Assessing the effects of random mutations in different genetic backgrounds

In Chapter 2 and 3, I studied whether and how the structural properties of the regulatory network were associated with the observed differences in gene expression over time. One way the regulatory interactions could influence the evolutionary process is that the genetic changes affecting those interactions could also change the properties of mutational effects. The idea that the existing genetic variants could affect mutational effects is represented by the concept of epistasis, and it is suggested that epistasis widely exist among different mutations (McKenzie et al. 1982; Remold and Lenski 2004; Milloz et al. 2008; Dworkin et al. 2009; Hansen 2013). Because the direction of evolution is partially determined by the explorable space depicted by the random mutations, epistasis could be important for phenotypic evolution by its role in reshaping the mutational effects based on pre-existing genetic variations. This idea is captured by multiple evolutionary studies, which illustrate that how evolution proceeds in the genetic level depends on existing genetic variations in the genome (Bridgham et al. 2006; Weinreich et al. 2006; Lang and Desai 2014). Thus, the knowledge on how standing genetic variants can influence properties of new mutations is necessary for answering the questions related to the predictability of the genetic basis of evolution.

In Chapter 4, I examined whether and how the mutational effects on the mean level of expression and the expression noise were dependent on existing genetic variants that

disrupt the regulation of gene expression. My results suggest that various aspects of mutational effects on gene expression are affected by the existence of pioneer genetic variants..

First, I showed that the *Saccharomyces cerevisiae* lab strain BY was less sensitive to random mutations on the mean level of expression, compared to strains carrying a genetic variant. This observation is consistent with the predictions from genetic canalization (Waddington 1942). Interestingly, mutational effects on the expression noise did not differ significantly among different starting genotypes (including the wild type).

I also found that the correlations between the mutational effects on the mean level of expression and the expression noise were dependent on the existing genetic variants. A central questions in the evolutionary studies is to predict the genetic basis of phenotypic evolution. However, one issue that prevents a straightforward prediction for the evolutionary trajectory is that mutations might have correlated effects on many traits. Intuitively speaking, if mutational effects on two traits are correlated, then selection might not be effective in driving the evolution of either trait, because genetic changes affecting one trait might affect the other trait simultaneously, in a way that might not be preferred by the selection on the other trait. This argument lays down the rationale behind which selection is less effective for suits of traits in natural population due to the existence of correlations between mutational effects on different traits (Pitchers et al. 2014). Considering the inconsistent theories and observations on evolutionary consequences of the expression noise (Fraser et al. 2004; Silander et al. 2012; Vardi et al. 2013; Zhang et al. 2009) and potential conflicts between the evolution of the mean level

of gene expression and the expression noise (Lehner 2008; Wolf et al. 2015), it is even harder to predict how multiple aspects on the gene expression could evolve together. One important question on this topic is that whether differing types of correlation between mutational effects on the mean level of expression and the expression noise could be set up to fit varying selection constraints imposed on the mean level of expression and the expression noise. Unlike what was found before (Thattai and Van Oudenaarden 2001; Munsky et al. 2012; Hornung et al. 2012; Murphy et al. 2010), our results suggested that prior genetic variants can induce either positive or negative correlation between the mutational effects on the mean level of expression and the expression noise. Thus, if the selection prefers opposite direction of changes in the expression noise and the mean level, then we would expect that the genetic variants that produce a negative correlation should be preferentially fixed in the population, and vice versa..

Finally, I found that the sensitivity to random mutations on the mean level of expression was positively correlated with the expression noise across different genetic variants in the promoter. It has been demonstrated that the cis-elements within a promoter is critical in determining the expression noise (Sanchez et al. 2013; Kim and Marioni 2013; Sharon et al. 2014). Thus, it is under expectation that different genetic variants in *cis* region can result in varying level of effects on the expression noise. However, the link between the expression noise and the mutational variance for different genetic variants in the promoter suggests that there could be a molecular mechanism that connects those two together. Also, because the sensitivity to random mutations is inverse proportional to the robustness to random mutations, my result also suggests that

increasing expression noise is associated with decreasing robustness to random mutations. Thus, if the robustness to random mutations is preferred by natural selection, then minimizing the expression noise might be favored by selection in this scenario, as predicted by multiple theoretical studies (Fraser et al. 2004; Silander et al. 2012).

Due to the scale of the experiments, only 8 genetic variants were analyzed in this experiment, and the number of mutants collected for each genetic background is limited (~300). The conclusions from Chapter 4 might be affected by the limited sample space explored in at least two ways. First, the descriptions of the mutational effects on the mean level of expression and the expression noise might be bias by the randomness in sorting out limited number of mutants. Second, the specific patterns or correlations I observed for the mutational effects in different genetic background might be due to the unique yet unknown properties of the genetic variants chosen in this study. Thus, to get a better idea how generalizable my findings are, more genetic backgrounds and more replicated are needed, as well as experimental systems in other species.

Taken together, my results demonstrate whether and how the existing genetic variants can influence various properties of mutational effects. Those observed patterns provide basis for a better understanding of the predictability of the genetic basis of the evolution in gene expression. However, in this study, I only focused on the overall distribution of mutational effects in different genetic backgrounds. In the future, detailed information on the characteristics of individual mutations collected in each genetic background could be obtained to get better insights on the relationship between properties of new mutations and existing genetic variations. For example, influence of new mutations on evolutionary trajectory is ultimately determined by their effects on chance

of survival of the organism, or fitness. Although I showed that different starting genetic variants can have different distributions of effect sizes from random mutations, it is not clear whether distributions of effects on fitness are different due to different starting expression level. Thus, to better understand how founding genetic variations could affect evolutionary fate of gene expression, it is necessary to measure the relationship between expression level and fitness of the organism.

Alon U. 2007. Network motifs: theory and experimental approaches. Nature Reviews Genetics 8:450–461.

Boley N, Wan KH, Bickel PJ, Celniker SE. 2014. Navigating and mining modENCODE data. Methods 68:38–47.

Borneman AR, Leigh-Bell JA, Yu H, Bertone P, Gerstein M, Snyder M. 2006. Target hub proteins serve as master regulators of development in yeast. Genes Dev. 20:435–448.

Bridgham JT, Carroll SM, Thornton JW. 2006. Evolution of Hormone-Receptor Complexity by Molecular Exploitation. Science 312:97–101.

Carreto L, Eiriz MF, Gomes AC, Pereira PM, Schuller D, Santos MAS. 2008. Comparative genomics of wild type yeast strains unveils important genome diversity. BMC Genomics 2014 15:1 9:524.

Carroll SB. 2008. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. Cell 134:25–36.

Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh C-H, Minokawa T, Amore G, Hinman V, Arenas-Mena C, et al. 2002. A Genomic Regulatory Network for Development. Science 295:1669–1678.

Dworkin I, Kennerly E, Tack D, Hutchinson J, Brown J, Mahaffey J, Gibson G. 2009. Genomic Consequences of Background Effects on scalloped Mutant Expressivity in the Wing of Drosophila melanogaster. Genetics 181:1065–1076.

Featherstone DE, Broadie K. 2002. Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network. Bioessays 24:267–274.

Fraser HB, Hirsh AE, Giaever G, Kumm J, Eisen MB. 2004. Noise minimization in eukaryotic gene expression.Ken Wolfe, editor. PLOS Biol 2:e137.

Goymer P. 2008. Network biology: why do we need hubs? Nature Reviews Genetics 9:650.

Hansen TF. 2013. WHY EPISTASIS IS IMPORTANT FOR SELECTION AND ADAPTATION. Evolution 67:3501–3511.

Hornung G, Bar-Ziv R, Rosin D, Tokuriki N, Tawfik DS, Oren M, Barkai N. 2012. Noise-mean relationship in mutated promoters. Genome Res. 22:2409–2417.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423:241–254.

Kim JK, Marioni JC. 2013. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. Genome biology.

Landry CR, Lemos B, Rifkin SA, Dickinson WJ, Hartl DL. 2007. Genetic Properties Influencing the Evolvability of Gene Expression. Science 317:118–121.

Lang GI, Desai MM. 2014. The spectrum of adaptive mutations in experimental evolution. Genomics 104:412–416.

Lehner B. 2008. Selection to minimise noise in living systems and its implications for the evolution of gene expression. Molecular Systems Biology 4:170.

Lu X, Jain VV, Finn PW, Perkins DL. 2007. Hubs in biological interaction networks exhibit low changes in expression in experimental asthma. In: Vol. 3. p. 98.

Macneil LT, Walhout AJM. 2011. Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. Genome Res. 21:645–657.

McGregor AP, Orgogozo V, Delon I, Zanet J, Srinivasan DG, Payre F, Stern DL. 2007. Morphological evolution through multiple cis-regulatory mutations at a single gene. Nature 448:587–590.

McKenzie JA, Whitten MJ, Adena MA. 1982. The effect of genetic background on the fitness of diazinon resistance genotypes of the Australian sheep blowfly, Lucilia cuprina. Heredity 49:1–9.

Milloz J, Duveau F, Nuez I, Félix M-A. 2008. Intraspecific evolution of the intercellular signaling network underlying a robust developmental system. Genes Dev. 22:3064–3075.

Munsky B, Neuert G, Van Oudenaarden A. 2012. Using gene expression noise to understand gene regulation. Science.

Murphy KF, Adams RM, Wang X, Balázsi G, Collins JJ. 2010. Tuning and controlling gene expression noise in synthetic gene networks. Nucl. Acids Res. 38:2712–2726.

Newman M., 2010. Network: An Introduction. Oxford University Press

Peisajovich SG. 2012. Evolutionary synthetic biology. ACS Synth Biol 1:199–210.

Pitchers W, Wolf JB, Tregenza T, Hunt J, Dworkin I. 2014. Evolutionary rates for multivariate traits: the role of selection and genetic variation. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 369:20130252–20130252.

Remold SK, Lenski RE. 2004. Pervasive joint influence of epistasis and plasticity on mutational effects in Escherichia coli. Nature Genetics 36:423–426.

Sanchez A, Choubey S, Kondev J. 2013. Regulation of noise in gene expression. Annual review of biophysics.

Sharon E, van Dijk D, Kalma Y, Keren L, Manor O, Yakhini Z, Segal E. 2014. Probing the effect of promoters on noise in gene expression using thousands of designed sequences. Genome Res. 24:1698–1706.

Silander OK, Nikolic N, Zaslaver A, Bren A, Kikoin I, Alon U, Ackermann M. 2012. A genome-wide analysis of promoter-mediated phenotypic noise in Escherichia coli.Matic I, editor. PLoS Genet. 8:e1002443.

Stern DL, Orgogozo V. 2008. THE LOCI OF EVOLUTION: HOW PREDICTABLE IS GENETIC EVOLUTION? Evolution 62:2155–2177.

Stern DL, Orgogozo V. 2009. Is Genetic Evolution Predictable? Science 323:746–751.

Thattai M, Van Oudenaarden A. 2001. Intrinsic noise in gene regulatory networks.

Vardi N, Levy S, Assaf M, Carmi M, Barkai N. 2013. Budding yeast escape commitment to the phosphate starvation program using gene expression noise. Curr. Biol. 23:2051–2057.

Vermeirssen V, Barrasa MI, Hidalgo CA, Babon JAB, Sequerra R, Doucette-Stamm L, Barabási A-L, Walhout AJM. 2007. Transcription factor modularity in a gene-centered C. elegans core neuronal protein-DNA interaction network. Genome Res. 17:1061–1071.

Waddington CH. 1942. Canalization of Development and the Inheritance of Acquired Characters. Nature 150:563–565.

Weinreich DM, Delaney NF, DePristo MA, Hartl DL. 2006. Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. Science 312:111–114.

Wolf L, Silander OK, van Nimwegen E. 2015. Expression noise facilitates the evolution of gene regulation. Elife 4:987.

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. Nature

Reviews Genetics 8:206–216.

Zhang Z, Qian W, Zhang J. 2009. Positive selection for elevated gene expression noise in yeast. Molecular Systems Biology 5:299.

Appendix

Assessing effects of mutations in different genetic locations

Reporter genes are widely used in studies on gene expression regulation. The system used in multiple mutagenesis experiments from Wittkopp Lab is the P_{TDH3}-YFP reporter construct located in *HO* locus (Duveau et al. 2014; Metzger et al. 2015; Metzger et al. 2016). However, it is not clear whether mutations affecting reporter expression level would have the same effect on native *TDH3* gene.

To examine this, I constructed a strain carrying no P_{TDH3}-YFP reporter construct in the *HO* locus. Instead, YFP coding sequence was inserted at the 3' end of the native *TDH3* coding sequence to produce a *TDH3:YFP* fusion protein at the native *TDH3* locus (strain YPW1452), so that expression level of native *TDH3* protein could be measured using flow cytometer. I then introduced 17 *cis* mutations, whose effects when inserted into P_{TDH3}-YFP in *HO* locus have been quantified previously (Metzger et al. 2015), into native *TDH3* promoter in the strain carrying the fusion protein. I then measured the expression level of all 18 genotypes (YPW1452 + 17 strains carrying both fusion protein and altered promoters in native *TDH3* locus) using flow cytometer and examined correlation with expression level measured from reporter construct in *HO* locus (Figure A.1). Both the mean level of expression and expression noise showed good correlation between reporter gene in *HO* locus and fusion protein in the native locus ($R^2 > 0.99$ for both cases). This result suggested that effects of mutations on reporter gene reflected

same effects on native protein expression, which provided supports on using the reporter gene to understand gene expression regulation in the *TDH3* system. This result was published in Metzger et al. (2016).

References

Duveau F, Metzger BPH, Gruber JD, Mack K, Sood N, Brooks TE, Wittkopp PJ. 2014. Mapping small effect mutations in Saccharomyces cerevisiae: impacts of experimental design and mutational properties. G3 (Bethesda) 4:1205–1216.

Metzger BPH, Duveau F, Yuan DC, Tryban S, Yang B, Wittkopp PJ. 2016. Contrasting Frequencies and Effects of cis- and trans-Regulatory Mutations Affecting Gene Expression. Mol. Biol. Evol. 33:1131–1146.

Metzger BPH, Yuan DC, Gruber JD, Duveau F, Wittkopp PJ. 2015. Selection on noise constrains variation in a eukaryotic promoter. Nature 521:344–347.

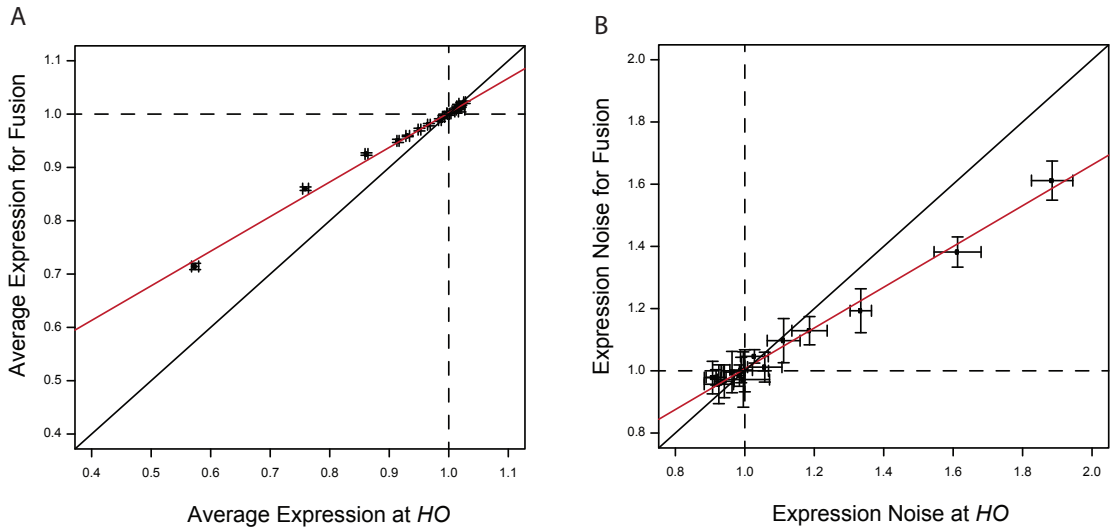

Figure A.1. Effects of 17 *cis* mutations on P_{TDH3} expression in different genomic location. Effects of individual *cis* mutations on the mean level of expression or noise of P_{TDH3}-*YFP* reporters integrated into the yeast genome at the *HO* locus or fused to the native TDH3 protein. Black solid line represents diagonal line. Dashed lines are the non-mutant control expression for each reporter. Red solid line is the slope from a linear regression. Error bars are 95% CI. (**A**) Effect of *cis*-regulatory mutations on the mean level of expression of P_{TDH3}-*YFP* reporter at the *HO* locus (x-axis) vs the effect of the same *cis*-regulatory mutations on the mean level of expression for the fusion protein in native *TDH3* locus. (**B**) Effect of *cis*-regulatory mutations on expression noise of P_{TDH3}-*YFP* reporter at the *HO* locus (x-axis) vs the effect of the same *cis*-regulatory mutations on expression noise for the fusion protein at the native *TDH3* locus.